Appunti di analisi dati

Studenti II anno Laurea Triennale in Fisica

A.A. 2021/2022

Indice

1 Lezione 1 - 11/10/2021

1.1 Metodo dei minimi quadrati per la stima dei parametri di una retta

Siano Y e X due grandezze per le quali si vuole verificare un andamento lineare $Y = mx + q$. Dato un campione costituito da n coppie (x_i, y_i) di valori misurati, con incertezze trascurabili su x_i e σ_i su y_i, secondo tale metodo la migliore stima di m e di q è quella che minimizza la funzione:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - y_i^t)^2}{\sigma_i^2}$$

dove $y_i^t = mx_i + q$. In altre parole, bisogna minimizzare lo scarto tra y_i e y_i^t. Si impongono allora le condizioni, in cui si svolgono le derivate rispetto a ciascun parametro da stimare e le si pongono pari a zero:

$$\begin{cases} \frac{\partial X^2}{\partial m} = 0 \\ \frac{\partial X^2}{\partial q} = 0 \end{cases} \qquad \text{da cui si ottengono le stime:}$$

$$\hat{m} = \frac{S_{00}S_{11} - S_{10}S_{01}}{D} \qquad \hat{q} = \frac{S_{01}S_{20} - S_{11}S_{10}}{D} \qquad \text{dove}$$

$$D = S_{00}S_{20} - S_{10}^2 \qquad S_{jk} = \sum_{i=1}^{n} \frac{x_i^j y_i^k}{\sigma_i^2}$$

e relative varianze ottenute dalla legge di propagazione della varianza:

$$\sigma_{\hat{m}}^2 = \frac{S_{00}}{D} \qquad e \qquad \sigma_{\hat{q}}^2 = \frac{S_{20}}{D}$$

$\hat{m}$ e $\hat{q}$ sono correlati, hanno una covarianza non nulla. Inoltre, abbiamo ottenuto tali stime considerando importanti le incertezza sulle y_i e trascurando quelle sulle x_i. Se fosse stato il contrario, sarebbe stato sufficiente scambiare il ruolo di X e di Y. Se, invece, entrambe le incertezze non fossero state trascurabili, il procedimento da adottare sarebbe leggermente più complesso:

1. All'ordine zero si ottengono le stime $\hat{m}_0$ e $\hat{q}_0$ assumendo che le incertezze sulle x_i siano trascurabili, il che consiste in un'approssimazione. In questo modo si ottiene un'idea sulla dipendenza tra X e Y.

2. Si trova come l'errore sulle x_i influiscono su quello delle y_i. Siccome:

   $$y_i = mx_i + q$$

   una σ_{x_i} mi darà una σ_{y_i}. In più c'è un'incertezza

   $$\sigma'_{y_i} = m\sigma_{x_i}$$

   data dalla legge di propagazione della varianza. Quindi, si definisce una varianza locale sulle y_i pari a:

   $$\sigma_{i,TOT}^2 = \sigma_i^2 \cdot \hat{m}_0^2 \sigma_{x_i}^2$$

   dove il primo fattore è la varianza sperimentale mentre il secondo è la varianza dovuta alle incertezze sulle x_i. A questo punto si stanno trascurando gli errori sulle x_i (perché abbiamo trasformato l'incertezza sulle x_i in un'incertezza aggiuntiva sulle y_i) e si possono applicare le formule trovate prima.

3. Si itera il ragionamento trovando i valori di $\hat{m}_i$ e $\hat{q}_i$ che forniscono la stima migliore.

Se le incertezze sulle x_i o sulle y_i sono simili tra loro, allora i coefficienti angolari trovati a vari ordini saranno anch'essi simili. Altrimenti, tali valori possono essere diversi.

Talvolta le leggi fisiche di cui si vogliono calcolare i parametri non sono lineari. Spesso però è possibile effettuare una linearizzazione. Per esempio:

$$V(t) = V_o e^{-\frac{t}{c}}$$

rappresenta la legge per la scarica di un condensatore, avente un andamento esponenziale e non lineare. Si applica il logaritmo naturale a entrambi i membri dell'equazione:

$$\ln V = \ln V_0 + \ln(e^{-\frac{t}{c}}) = \ln V_0 - \frac{t}{c}$$

Se si impone $y = \ln V$ e $x = t$, si ritrova una relazione di tipo lineare:

$$y = \ln V_0 - \frac{x}{c}$$

I parametri stimati, essendo funzioni di variabili casuali, sono anch'essi delle variabili casuali. Dunque, hanno una funzione di distribuzione. L'errore statistico associato assume un significato differente a seconda della funzione di distribuzione che descrive i parametri stimati.

In laboratorio, normalmente, dominano gli errori dovuti alla sensibilità degli strumenti e non quelli statistici. Però, si assume che la distribuzione sia uniforme entro gli errori massimi. Dunque, si usa come errore statistico la deviazione standard della distribuzione. Ricordiamo che la distribuzione uniforme è così definita:

$$x \in (a,b), f(x) = \begin{cases} \frac{1}{c} & a < x < b \\ 0 & x < a \vee x > b \end{cases}$$

dove $c = b - a$. La varianza associata è:

$$\sigma_x^2 = \frac{(b-a)^2}{12} = \frac{(2\Delta x)^2}{12} = \frac{\Delta x^2}{3}$$

da cui si può ricavare la relazione di conversione tra errori massimi e statistici:

$$\sigma_x = \frac{\Delta x}{\sqrt{3}}$$

2 Variabili casuali e distribuzioni di probabilità

In questo corso ci occuperemo di fenomeni statistici, caratterizzati dalle cosiddette "variabili casuali" e nei quali trascuriamo i contributi dati dagli errori massimi e da quelli sistematici. In altre parole, ci limitiamo a trattare le fonti di incertezza di natura statistica. Tali aspetti potrebbero manifestarsi in particolari situazioni oppure risultare intrinsechi al fenomeno considerato. Per esempio, quando non è possibile misurare direttamente il valore vero di una grandezza.

Vediamo ora alcuni esempi di variabili casuali di natura statistica:

1. Fenomeni di diffusione (*scattering*): coinvolgono due particelle che urtano tra loro e il cui angolo di diffusione non ha sempre lo stesso valore, ogni intervallo di valori ha una probabilità che dipende dalla sezione d'urto. Si può cercare di risalire alla fisica che governa questo processo ripetendo la misura più volte e studiando la distribuzione.

2. Periodo di oscillazione di un pendolo semplice: ponendosi in questa approssimazione per piccoli angoli, data la lunghezza ed il periodo di oscillazione (che è soggetta a errori accidentali importanti) si può calcolare l'accelerazione di gravità. Si ripete n volte le misure, identificando il campione $t_1, \ldots, t_n$, ovvero un insieme di n variabili casuali indipendenti. Di conseguenza, anche il campione stesso è casuale perché se si ripetono le n misure, si ottiene un campione differente. La stima del periodo di oscillazione è:

$$\hat{T} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

e la varianza associata:

$$\sigma_{\hat{T}} = \frac{1}{n-1}\sum_{i}(t_i - \hat{T})^2$$

Al campione viene associata una certa funzione di distribuzione. Inoltre, sia $\hat{T}$ che $\sigma^2_{\hat{T}}$ sono a loro volta delle variabili casuali perché funzioni di variabili casuali.

Riassumendo, dobbiamo essere in grado di descrivere fenomeni nei quali intervenga sia una che più variabili casuali e riuscire a lavorare con funzioni di variabili casuali.

2.1 Variabili casuali discrete

Una variabile X è discreta se può assumere valori su un insieme finito o numerabile di valori: $\Omega_X = \{x_1, \ldots, x_n\}$, ciascuno con una probabilità: $\{p_1, \ldots, p_n\}$, con $p_i = P(X = x_i)$, la successione delle p_i forma la distribuzione di probabilità. Si ha $p_i \geq 0, \sum_{i=1}^{n} p_i = 1$. Data una funzione di X, il valore di aspettazione è

Definizione 1 (Valore atteso per variabili discrete)**.**

$$E[f(x)] = \sum_i f(x_i)p_i.$$

Il valore di aspettazione di X è

$$\mu = E[X] = \sum_i x_i p_i$$

ed è un indice di posizione. Un altro valore di aspettazione importante è

$$E[(X-\mu)^2] = \sum_i (x_i - \mu)^2 p_i = \sigma_X^2$$

La sua radice quadrata è la deviazione standard: σ_X ed è un indice di dispersione.

Nota:

$$\sigma_X^2 = \sum_i x_i^2 p_i + \sum_i \mu^2 p_i - \sum_i 2\mu x_i p_i = E[X^2] + \mu^2 - 2\mu E[X] = E[X^2] - (E[X])^2.$$

2.2 Variabili casuali continue

Una variabile casuale X è continua se assume valori all'interno di un intervallo: $\Omega_X = (a, b)$. Si noti che $P(X = x^*) = 0$. Si può però definire $f(x)$ funzione di distribuzione, con

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)dx$$

che risulterà positiva, limitata e normalizzata. La condizione di normalizzazione diventa

$$\int_{\Omega_X} f(x)dx = 1$$

Il valore di aspettazione diventa

$$E[X] = \int_{\Omega_X} xf(x)dx = \mu.$$

Analogamente

$$\text{var}(X) = \sigma_X^2 = \int_{\Omega_X} (x-\mu)^2 f(x)dx.$$

Si potrebbero definire altri indici di dispersione, come ad esempio $E[|x-\mu|]$, ma non godrebbero di alcune proprietà molto utili della varianza, quale la seguente:

Teorema 1 (disuguaglianza di Čebyšëv)**.**

$$P(|x-\mu| \geq \lambda\sigma) \leq \frac{1}{\lambda^2}.$$

Dimostrazione. Data una funzione positiva h di X, positiva su tutto Ω_X, se vale $h(x) \geq k, \forall x \in R$ con $R \subseteq \Omega_X$ intervallo, allora

$$E[h(X)] = \int_{\Omega_X} h(x)f(x)dx \geq \int_R h(x)f(x)dx \geq k\int_R f(x)dx = kP(x \in R),$$

cioè $E[h(X)] \geq kP(h(X) \geq k)$. Ponendo $h(X) = (X-\mu)^2, k = (\lambda\sigma)^2$, si ottiene la disuguaglianza di Čebyšëv: $E[(X-\mu)^2] = \lambda^2\sigma^2 P((X-\mu)^2 \geq \lambda^2\sigma^2)$, da cui $\sigma^2 \geq \lambda^2\sigma^2 P(|x-\mu| \geq \lambda\sigma)$. □

Per questa disuguaglianza, si ha $P(|x-\mu| < 2\sigma) > 1 - 0.25 = 0.75$. Sappiamo che per una distribuzione gaussiana vale $P(|x-\mu| < 2\sigma) \approx 0.95$.

Definizione 2 (Distribuzione cumulativa)**.**

$$F(X) := \int_a^X f(X')dX'.$$

Risulta utile per generare dei numeri casuali che seguano una distribuzione data: se X è una variabile casuale con distribuzione uniforme in $[0,1]$, allora $y = F^{-1}(X)$ è una variabile casuale e si può dimostrare che segue la distribuzione $f(y)$.

I momenti di distribuzioni sono degli indici ulteriori che consentono una completa descrizione di una funzione di distribuzione di probabilità, li vedremo più avanti.

2.3 Distribuzione binomiale

Data una variabile che può avere come esiti/valori solo successo o insuccesso al verificarsi di ogni evento, sia p la probabilità di successo e $q = 1-p$ la probabilità di insuccesso. Supponiamo di ripetere l'esperimento N volte. Qual è la probabilità che ci siano k successi su N ripetizioni? La risposta è fornita dalla distribuzione binomiale, che ha la seguente forma:

$$p_k = \binom{N}{k} p^k q^{N-k}, \qquad \text{con} \quad k = 0, \ldots, N$$

Il valore di aspettazione è $E[k] = \sum_k kp_k = Np$, mentre la varianza vale $\sigma_X^2 = Np(1-p)$. Se si costruisce un istogramma di valori misurati che seguono una certa distribuzione, il numero di eventi che cadono in un singolo intervallo I_i segue, in prima approssimazione, la distribuzione di Bernoulli (se il numero di intervalli è sufficientemente elevato), quindi vale: $E[n_i] = Npi$, dove la probabilità p_i è data da $\int_{I_i} f(x)dx$.

2.4 Distribuzione di Poisson

Si ottiene dalla distribuzione binomiale per $N \to \infty, p \to 0$, con $Np = \nu$ costante. La distribuzione è

$$p_k = \frac{\nu^k e^{-\nu}}{k!},$$

con $\nu = E[k] = Np$. Vale $\sigma^2 = \nu$.

2.5 Momenti

Il valore di aspettazione μ e la varianza σ non sono sufficienti per descrivere compiutamente una funzione di distribuzione. Infatti, in realtà una funzione di distribuzione è caratterizzata da un numero infinito di momenti, che si distinguono in momenti algebrici e centrali, con relativo ordine:

Definizione 3 (Momento algebrico di ordine k)**.**

$$\mu_k^* = E[x^k] = \int_{\Omega_X} x^k f(x)dx$$

dove Ω_X è il dominio della variabile casuale X.

Ordine	Momento algebrico
0	$\mu_0^* = 1$
1	$\mu_1^* = \mu_x$
2	$\mu_2^* = \int_{\Omega_X} x^2 f(x)dx$ (difficile da interpretare)

Tabella 1: Momenti algebrici

In generale, tutti i momenti algebrici di ordine successivi a quelli indicati non forniscono particolari informazioni né caratterizzano in modo evidente la funzione di distribuzione.

Definizione 4 (Momento centrale di ordine k)**.**

$$\mu_k = E[(x - \mu_x)^k].$$

Questi momenti sono detti "centrali" perché sono riferiti a μ_x e non più all'origine. Così facendo, si ottengono maggiori informazioni sulla funzione di distribuzione. Al contrario dei momenti algebrici, tutti i momenti centrali, anche di ordine superiore al secondo, sono interessanti perché danno tutti maggiori informazioni sulla forma della funzione di distribuzione. Spesso si usano al posto di μ_3 e μ_4 il coefficiente di asimmetria $\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$ e il coefficiente di curtosi $\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$. γ_1 vale 0 se la distribuzione è simmetrica rispetto al valore di aspettazione (come nel caso della funzione di distribuzione di Gauss) e γ_2 vale 0 per la distribuzione di Gauss (grazie alla correzione di tre unità).

Ordine	Momento centrale	Nome
0	$\mu_0 = 1$	-
1	$\mu_1 = 0$	-
2	$\mu_2 = \int_{\Omega_X}(x-\mu_x)^2 f(x)dx = \sigma_x^2$	Varianza (indice di posizione)
3	$\mu_3 = E[(x-\mu_x)^3]$	Indice di assimmetria/distorsione/skewness
4	$\mu_4 = E[(x-\mu_x)^4]$	Indice di Curtosi (quanto è "piatta" f?)

Tabella 2: Momenti centrali

2.6 Distribuzione esponenziale

Una funzione di distribuzione esponenziale è nella forma

$$g(x) = \frac{1}{\tau}e^{-x/\tau}, x \geq 0.$$

e descrive, per esempio, la funzione di distribuzione dei tempi di attesa per eventi casuali indipendenti, come il decadimento di nuclei radioattivi.

È normalizzata:

$$\int_0^\infty g(x)dx = \frac{1}{\tau}\int_0^\infty e^{-x/\tau} = -e^{-x/\tau}\Big|_0^\infty = 1.$$

Il valore di aspettazione è:

$$\mu_x = E[x] = \int_0^\infty x\frac{1}{\tau}e^{-x/\tau} = xe^{-x/\tau}\Big|_0^\infty + \int_0^\infty e^{-x/\tau}dx = -\tau e^{-x/\tau}\Big|_0^\infty = \tau.$$

La varianza è:

$$\text{var}(X) = E[x^2] - E[x]^2 = \tau^2,$$

con

$$E[x^2] = \int_0^\infty x^2\frac{1}{\tau}e^{-x/\tau}dx = -x^2e^{-x/\tau}\Big|_0^\infty + \int_0^\infty 2xe^{-x/\tau}dx = 0 + 2\tau E[x] = 2\tau^2$$

Quindi sia il concetto di valore di aspettazione che di varianza sono legati a τ, caratteristica condivisa dalla funzione di distribuzione gaussiana. Tuttavia, quella esponenziale rimane molto diversa da quest'ultima perché ciò che le differenzia sono i momenti di ordine superiore. Inoltre, cambia anche il significato della deviazione standard: Di conseguenza, le due funzioni presentano

Distribuzione	$P(\mu_x - \sigma \leq X \leq \mu_x + \sigma)$
Gaussiana	0.68
Esponenziale	0.86 $(0 \leq x \leq 2\tau)$

Tabella 3: Confronto di significati della dev. std.

contenuti di probabilità diversi a parità di intervallo di partenza considerato. Si può notare che per la funzione di distribuzione gaussiana si possono considerare anche gli intervalli individuati da due o tre deviazioni standard. Ciò non può essere fatto per la distribuzione esponenziale perché si uscirebbe dall'intervallo di definizione della variabile casuale.

Ora torniamo al calcolo dei momenti...Per quanto riguarda il momento di asimmetria

$$\begin{aligned}\mu_3 &= E[(X-\mu_x)^3] = E[(X^3 - \mu_x^3 - 3X^2\mu_x + 3X\mu_x^2)] = \\ &= E[X^3] - \mu_x^3 - 3\mu_x E[x^2] + 3\mu_x^2 E[X] = \\ &= E[X^3] - \mu_x^3 - 6\mu_x\tau^2 + 3\mu_x^2\tau = \\ &= E[X^3] - 4\tau^3 = \int_0^\infty x^3 \frac{1}{\tau} e^{-x/\tau} dx - 4\tau^3 = \\ &= -x^3 e^{-x/\tau}\Big|_0^\infty + \int_0^\infty 3x^2 e^{-x/\tau} dx - 4\tau^3 = \\ &= 0 + 3\tau E[X^2] - 4\tau^3 = \\ &= 2\tau^3.\end{aligned}$$

Di conseguenza, $\gamma_1 = 2$. Vale anche $\mu_4 = 6\tau^4$, mentre per la distribuzione di Gauss vale $\mu_4 = 3\sigma^4$.

2.7 Funzioni generatrici dei momenti

La funzione generatrice dei momenti algebrici contiene tutte le informazioni sui momenti algebrici ed è definita come

$$M_x^*(t) = E[e^{tx}] = \int_{\Omega_X} e^{tx} f(x)dx = \sum_k \frac{1}{k!} \left\{ \int_{\Omega_X} x^k f(x)dx \right\} t^k = \sum_k \frac{1}{k!} \mu_k^* t^k,$$

da cui $\mu_k^* = \frac{d^k}{dt^k} M_x^*(t)\Big|_{t=0}$, che evidenzia proprio come sia possibile ottenere tutti i momenti a partire dalla funzione generatrice, tramite delle derivate.

Analogamente la funzione generatrice dei momenti centrali

$$M_x(t) = E[e^{t(x-\mu)}] = \int_{\Omega_X} e^{t(x-\mu)} f(x)dx = \sum_k \frac{1}{k!} \mu_k t^k,$$

da cui $\mu_k = \frac{d^k}{dt^k} M_x(t)\Big|_{t=0}$.

Osservazione 1 (Relazione tra funzioni generatrici dei momenti)**.**

$$M_x(t) = e^{-\mu t} M_x^*(t).$$

Nota: calcolare i momenti attraverso le derivate delle funzioni generatrici oppure applicando la definizione sono due metodi assolutamente equivalenti.

2.7.1 Distribuzione normale

Ad esempio, per una distribuzione normale si ha

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

e

$$M_x^*(t) = \int_{-\infty}^{+\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{-(x-\mu)^2/2\sigma^2 + tx} dx$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{-[(x-\mu)^2 - tx2\sigma^2]\frac{1}{2\sigma^2}} dx$$

Sviluppando i calcoli e applicando la sostituzione $y = (x - \mu - \sigma^2 t)$

$$-[(x-\mu)^2 - tx2\sigma^2]\frac{1}{2\sigma^2} = -\frac{1}{2\sigma^2}[(x-\mu-\sigma^2 t)^2 - \sigma^4 t^2 - 2\sigma^2\mu t] = -\left(\frac{y^2}{2\sigma^2} - \frac{\sigma^2 t^2}{2} - \mu t\right)$$

Nell'integrale otteniamo

$$M_x^*(t) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{-\frac{y^2}{2\sigma^2}} e^{\frac{\sigma^2 t^2}{2} + \mu t} dy = e^{\frac{\sigma^2 t^2}{2} + \mu t}$$

$$M_x(t) = e^{-\mu t} M_x^*(t) = e^{\sigma^2 \frac{t^2}{2}}$$

Si possono ora calcolare i momenti con le derivate. I calcoli per esercizio.

2.7.2 Distribuzione binomiale

$$p_k = \binom{n}{k} p^k q^{n-k}$$

$$M_k^*(t) = E[e^k t] = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} e^{kt} = \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k q^{n-k} = (pe^t + q)^n.$$

$$M_k(t) = (pe^t + q)^n e^{-tnp} = (pe^t e^{-tp} + qe^{-tp})^n = (pe^{tq} + qe^{-tp})^n$$

In particolare, si può calcolare il coefficiente di asimmetria:

$$\gamma_1 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{npq(q-p)}{(npq)^{\frac{3}{2}}} = \frac{npq(q-p)}{npq\sqrt{npq}} = \frac{q-p}{\sqrt{npq}}$$

da cui si osserva che esso è nullo quando $p = q = 0.5$. Tuttavia, a seconda della diversità di p e di q la funzione di distribuzione binomiale può essere molto asimmetrica, da una parte oppure dall'altra a seconda del segno del coefficiente di asimmetria. Se si considera:

$$\lim_n \gamma_1 = \lim_n \frac{q-p}{\sqrt{npq}} = 0$$

cioè se si ha un numero molto alto di eventi, la distribuzione è simmetrica, indipendentemente dai valori assunti da p e da q. Similmente, si può calcolare il coefficiente di curtosi:

$$\gamma_2 = \frac{1 - 6pq}{npq}$$

che in generale è diverso da zero qualunque siano i valori assunti da p e da q. Tuttavia, se si considera il limite:

$$\lim_n \gamma_2 = \lim_n \frac{1 - 6pq}{npq} = 0$$

ovvero la funzione è "piatta" come la funzione di distribuzione gaussiana. Riassumendo, quando si ha a disposizione un numero elevato di eventi, la funzione di distribuzione binomiale assume la stessa forma di una distribuzione normale $N(np, npq)$. Ne consegue che anche il contenuto di probabilità nei singoli intervalli è simile per le due distribuzioni:

$$P(\mu - \sigma \leq k \leq \mu + \sigma) = P(np - \sqrt{npq} \leq k \leq np + \sqrt{npq}) \approx P(\mu - \sigma \leq x \leq \mu + \sigma) = 0.68$$

2.7.3 Distribuzione esponenziale

$$f(x) = \frac{1}{\tau} e^{-x/\tau}$$

$$M_x^*(t) = \frac{1}{1 - t\tau}$$

$$M_x(t) = \frac{e^{-\tau t}}{1 - t\tau}$$

Teorema 2 (Caratterizzazione tramite momenti)**.** $f_1(x) = \sum_k a_k x^k \frac{1}{k!}$ $f_2(x) = \sum_k b_k x^k \frac{1}{k!}$ *Con momenti rispettivamente* $\mu_{k,1}^* = \mu_{k,2}^*$ *(coincidono tutti i momenti algebrici). Allora le due funzioni coincidono:*

Dimostrazione.

$$\int_{\Omega_X} (f_1 - f_2)^2 dx = \sum_k \frac{1}{k!}(a_k - b_k) \int_{\Omega_X} (f_1 - f_2) x^k dx =$$

$$= \sum_k \frac{1}{k!}(a_k - b_k) \Big\{ \underbrace{\int_{\Omega_X} x^k f_1(x) dx}_{\mu_{k,1}^*} - \underbrace{\int_{\Omega_X} x^k f_2(x) dx}_{\mu_{k,2}^*} \Big\} = 0.$$

Quindi è vero. □

Esempio: supponiamo di avere una variabile casuale $y = y(x)$ e la pdf $f(x)$, ma non $h(y)$, si ha

$$M_y^*(t) = E[e^{ty}] = \int_{\Omega_y} e^{ty} h(y) dy = \int_{\Omega_X} e^{ty(x)} f(x) dx$$

Se per caso si trova che:

$$M_y^* = e^{at + \frac{t^2}{b}}$$

allora sappiamo che y segue una funzione di distribuzione gaussiana. Infatti, la funzione generatrice dei momenti algebrici della distribuzione gaussiana è:

$$M_x^* = e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

allora per il Teorema 2 si ha che la distribuzione di y è determinata perché ha gli stessi valori di aspettazione della distribuzione normale $N(a, \frac{2}{b})$.

2.7.4 Distribuzione di Poisson

La distribuzione è $p_k = \frac{\nu^k}{k!} e^{-\nu}$

$$E[k] = \sum_k k p_k = \nu.$$

$$\mathrm{var}(k) = \sum_k (k - \nu)^2 p_k = \nu$$

$$M_k^*(t) = E[e^{tk}] = \sum_k e^{tk} \frac{\nu^k}{k!} e^{-\nu} = e^{-\nu} \sum_k \frac{1}{k!} (\nu e^t)^k = e^{-\nu} e^{\nu e^t} =$$

$$= e^{\nu(e^t - 1)}.$$

	$\nu = 0.1$	$\nu = 0.5$	$\nu = 5$
$k = 0$	0.90	0.61	0.07
$k = 1$	0.09	0.30	
$k = 2$	0.005	0.02	
$k = 3$	0.005	0.07	0.14
$k = 5$			0.17
$k = 7$			0.10

$$M_k(t) = E[e^{t(k-\nu)}] = e^{-\nu t} M_k^*(t) = e^{\nu(e^t - 1 - t)}.$$

Si possono fare calcoli anche qua. I risultati in Tabella 8. Esempio: un fenomeno statico produce eventi casuali indipendenti. Per n volte: conto quanti eventi capitano in un intervallo di tempo fissato Δt (chiamo ν il numero di eventi medio in Δt). Ottengo i conteggi: $k_1, \ldots, k_m$. Ogni valore k_j lo registro n_j volte. La variabile casuale discreta k_j segue in buona approssimazione Poisson, ma anche la variabile casuale n_j se n è grande e le p_j sono piccole, quindi vale che:

$$n_j^c := E[n_j] = np_j = n\frac{\nu^{k_j}}{k_j!}e^{-\nu}.$$

Siccome n_j segue Poisson:

$$\text{var}(n_j) = np_j = n_j^c$$

Per approfondimenti sulla distribuzione dei conteggi nelle colonne di istogrammi, andare a vedere nella sezione 2.14 la distribuzione multinomiale.

Esempio:

Un negozio ha in media 7 clienti all'ora. Qual è la probabilità che ne entrino 0 in mezz'ora? E 3?

7 clienti/ora=3.5 clienti/mezzora $\implies \nu = 3.5$

$p_0 = e^{-3.5} \approx 0.03$

$p_3 = \frac{3.5^3 e^{-3.5}}{6} \approx 0.22$

Esempio:

Una fabbrica fa cose che si rompono per fare una macchina. Una macchina ne ha 2000, la probabilità che una roba si rompa in un mese è $p = 0.0005 = 5 \cdot 10^{-4}$. $\nu = 5 \cdot 10^{-4} \cdot 2 \cdot 10^3 = 1$. Qual è la probabilità che la macchina si rompa in un mese?

$$p_0 = \frac{\nu^0 e^{-1}}{000!} \approx 0.37.$$

$$p_{rotta} = p(n_{guasti} \geq 1) = 1 - p(n_g = 0) \approx 0.63.$$

2.8 Distribuzione Gamma

$$f(x, \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta},$$

con

$$\Gamma(\alpha) = \int_0^\infty z^{\alpha-1} e^{-z} dz.$$

Ad esempio

$$\Gamma(1/2) = \int_0^\infty z^{1/2} e^{-z} dz \underset{dz=ydy}{\overset{y^2=2z}{=}} \int_0^\infty dy\sqrt{2} y^{-1} e^{-\frac{y^2}{2}} y = \frac{\sqrt{2}}{2}\sqrt{2\pi} = \sqrt{\pi}.$$

La distribuzione è normalizzata:

$$\int_0^\infty \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} dx = \frac{1}{\Gamma(\alpha)} \int_0^\infty \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-x/\beta} \frac{dx}{\beta} = 1.$$

Il valore di aspettazione è:

$$E[x] = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^\alpha e^{-x/\beta} dx = \frac{\beta}{\Gamma(\alpha)} \underbrace{\int_0^\infty \left(\frac{x}{\beta}\right)^{\alpha} e^{-x/\beta} \frac{dx}{\beta}}_{\Gamma(\alpha+1)} =$$

$$= \beta \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \alpha\beta.$$

Per ottenere la varianza calcoliamo

$$E[x^2] = \frac{\beta^2}{\Gamma(\alpha)} \int_0^\infty \frac{x^{\alpha+1}}{\beta^{\alpha+1}} e^{-x/\beta} \frac{dx}{\beta} = \frac{(\alpha+1)\alpha\Gamma(\alpha)\beta^2}{\Gamma(\alpha)} = (\alpha+1)\alpha\beta^2$$

Da cui

$$\text{var}(x) = E[x^2] - (E[x])^2 = (\alpha+1)\alpha\beta^2 - \alpha^2\beta^2 = \alpha\beta^2.$$

2.9 Distribuzione di Erlang

(Caso particolare della funzione di distribuzione Γ con: $\alpha = k \in \mathbb{N}^+ b$ e $\beta = 1$)

$$f(x,k) = \frac{1}{(k-1)!} x^{k-1} e^{-x}$$

$$E[x] = k, \sigma_k^2 = k.$$

Teorema 3 (Funzione di distribuzione di tempi di attesa)**.** *La funzione di distribuzione del tempo d'attesa per il k-esimo di un insieme di eventi casuali, equiprobabili e indipendenti tra di loro, è una funzione di Erlang $f(x,k)$, rispetto a $x = \mu t = t/\tau$ e $\mu = 1/\tau$ il numero medio di eventi per unità di tempo.*

Dimostrazione. Posso calcolare la probabilità che l'evento *k-esimo* cada in un intervallo di tempo $(t, t+\delta t)$. Essa è data dalla probabilità di aver misurato il $(k-1)$*-esimo* evento nell'intervallo $(0,t)$ per quella di misurare un evento nell' intervallo $(t, t+\delta t)$. Se

- μ è il numero medio di eventi per unità di tempo
- e δt è sufficientemente piccolo da non poter contenere più di un evento (è impossibile avere eventi contemporanei)

allora la probabilità che *un* evento si verifichi nell'intervallo $(t, t+\delta t)$ è:

$$p(t, t+\delta t) = \mu\delta t$$

e dunque, siccome gli eventi sono indipendenti, la probabilità che *il k-esimo* evento si sia verificato nell'intervallo $(t, t+\delta t)$ è:

$$p_k(t, t+\delta t) = p_{k-1}(0,t) \cdot p(t, t+\delta t) = p_{k-1}(0,t)\mu\delta t$$

dove $p_{k-1}(0,t) = \frac{\nu^{k-1}}{(k-1)!}e^{-\nu}$ (con ν numero medio di eventi in (0,t) $\nu = \mu t$)[1] allora

$$p_k(t, t+\delta t) = \frac{(\mu t)^{k-1}}{(k-1)!}e^{-\mu t}\mu\delta t$$

allora la funzione di distribuzione dei tempi d'attesa del k-esimo evento è

$$f_k(t) = \frac{\mu^k t^{k-1}}{(k-1)!}e^{-\mu t}.$$

Se pongo $x = \mu t$ nella funzione di Erlang, trovo stessa espressione (detta $g_k(x)$ la distribuzione di Erlang con il parametro k fissato, $f_k(t)$ la distribuzione del tempo di attesa del k-esimo evento: $f_k(t) = g_k(x)|\frac{dx}{dt}|$) □

$k=1$	$f_1(t) = \mu e^{-\mu t} = \frac{1}{\tau}e^{-t/\tau}$	$E[t] = \tau$
$k=2$	$f_2(t) = \frac{1}{\tau^2}te^{-t/\tau}$	$E[t] = 2\tau$
$k=3$	$f_3(t) = \frac{1}{2\tau^3}t^2e^{-t/\tau}$	

Tabella 4: Funzioni di distribuzione del primo, secondo e terzo evento

2.10 Funzione di distribuzione di χ^2

(Caso particolare finzione di distribuzione di Γ con $\alpha = n/2$ e $\beta = 2$)

$$f_n(x) = \frac{x^{\frac{n}{2}-1}e^{-x/2}}{2^{n/2}\Gamma(\frac{n}{2})}$$

$$E[x] = n, \qquad \sigma_x^2 = 2n$$

Esempi:

$n=1$	$f_1(x) = \frac{1}{\sqrt{2\pi}}x^{-1/2}e^{-x/2}$
$n=2$	$f_2(x) = \frac{1}{2}e^{-x/2}$
$n=3$	$f_3(x) = \frac{1}{2^{3/2}\sqrt{\pi}/2}x^{1/2}e^{-x/2}$
$n=4$	$f_3(x) = \frac{1}{4}xe^{-x/2}$

Tabella 5: Funzioni di distribuzione del χ^2 con n gradi di libertà

Le sue funzioni generatrici dei momenti sono:

$$M_x^*(t) = E[e^{tx}] = (1-2t)^{-n/2}$$

$$M_x(t) = e^{-tn}(1-2t)^{-n/2}$$

Possiamo allora calcolare i primi momenti:
allora

· $\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{8}{2\sqrt{2n}}$ (tende a 0 per n che tende a infinito)

[1] si tratta della funzione di distribuzione di Poisson, che regola i conteggi in un dato intervallo di tempo

$\mu_1^* = n$	$\mu_1 = 0$
	$\mu_2 = 2n$
	$\mu_3 = 8n$
	$\mu_4 = 12n^2 + 48n$

Tabella 6: Alcuni momenti della funz. di distribuzione del χ^2

· $\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{12}{n}$ (tende a 0 per n che tende a infinito)

Teorema 4. *Se n tende a ∞, la distribuzione χ_n^2 tende a $N(n, 2n)$. Ponendo $y = \frac{x-n}{\sqrt{2n}} = \frac{x-\mu_x}{\sigma_x}$ si ottiene $N(0, 1)$.*

Dimostrazione. Basta dimostrare che la funzione generatrice dei momenti algebrici delle due distribuzioni coincide per $n \to \infty$. Calcolo dunque $M_y^*(t)$:

$$M_y^*(t) = E[e^{ty}] = E[e^{\frac{tx}{\sqrt{2n}}} e^{-t\sqrt{\frac{n}{2}}}] = E[e^{\frac{tx}{\sqrt{2n}}}] \cdot e^{-t\sqrt{\frac{n}{2}}}$$
$$= M_x^*\left(\frac{t}{\sqrt{2n}}\right) \cdot e^{-t\sqrt{\frac{n}{2}}} = \left(1 - \frac{2t}{\sqrt{2n}}\right)^{-n/2} \cdot e^{-t\sqrt{\frac{n}{2}}}$$

Ne calcolo il logaritmo, usando lo sviluppo in serie di Taylor centrato in $x = 1$, ossia $\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} \ldots$

$$\ln M_y^* = -t\sqrt{\frac{n}{2}} - \frac{n}{2} \ln\left(1 - t\sqrt{\frac{2}{n}}\right)$$
$$= -t\sqrt{\frac{n}{2}} - \frac{n}{2}\left[- t\sqrt{\frac{2}{n}} - \frac{t^2}{n} - \frac{t^3}{3}\left(\frac{2}{n}\right)^{3/2} \ldots \right] = \frac{t^2}{2} + \frac{\sqrt{2}}{3}\frac{t^3}{\sqrt{n}} + \cdots \to \frac{t^2}{2}$$

dunque $\lim_{n\to\infty} M_y^* = e^{t^2/2}$ che corrisponde alla funzione generatrice dei momenti algebrici per la distribuzione normale standard. □

2.11 Distribuzione di Maxwell-Boltzmann

Si dimostra (cfr. Quando lo faremo) che una variabile con distribuzione di χ^2 è la somma dei quadrati di variabili con distribuzione normale standard (se sono indipendenti).

Da questo risultato si ricava la distribuzione di Maxwell-Boltzmann per le velocità delle molecole di un gas perfetto. Si supponga che:

1. Le particelle si muovano in direzione casuale
2. Non ci sia una direzione privilegiata
3. Non ci sia interazione tra le molecole

Le particelle si muovono con una velocità $v = |\vec{v}|$, la cui funzione di distribuzione vogliamo ricavare. Fissato un sistema di riferimento cartesiano, Sia $\vec{v} = (v_x, v_y, v_z)$, e supponiamo per semplicità che v_i abbia distribuzione normale $\forall i \in \{x, y, z\}$ (di solito è quella più semplice da supporre quando ci sono numerosi effetti casuali di cui non si conosce l' origine). Siccome non ci sono direzioni privilegiate, esse risultano indipendenti e centrate in 0. Allora

$$f(v_i) = N(0, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-v_i^2/2\sigma^2}.$$

Introduciamo $\vec{q} = \vec{v}/\sigma$ (vettore che ha la stessa direzione di v), $q_i = v_i/\sigma$. Per calcolare la distribuzione di q_i bisogna mantenere invariata la probabilità elementare $g(q_i)dq_i = f(v_i)dv_i$, da cui $g(q_i) = f(v_i)\left|\frac{dv_i}{dq_i}\right|_{q_i}$. Vale quindi

$$g(q_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\sigma^2 q_i^2/2\sigma^2}\sigma = \frac{1}{\sqrt{2\pi}} e^{-q_i^2/2} = N(0,1).$$

Sia ora $q^2 = q_x^2 + q_y^2 + q_z^2 =: z = \frac{v^2}{\sigma^2}$ (a noi infatti ci interessa il modulo di v). Siccome q_i hanno distribuzioni normali standard, dal risultato menzionato all'inizio si ha che

$$g(z) = \frac{\sqrt{z}e^{-z/2}}{2^{3/2}\Gamma(3/2)}$$

ma $v = \sigma\sqrt{z}$, quindi

$$h(v) = g(z(v))\left|\frac{dz}{dv}\right| = \frac{1}{2\sqrt{2}\frac{\sqrt{\pi}}{2}}\frac{v}{\sigma}e^{-v^2/2\sigma^2}\frac{2v}{\sigma^2} = \sqrt{\frac{2}{\pi}}\frac{v^2}{\sigma^3}e^{-v^2/2\sigma^2}.$$

Ponendo $\sigma^2 = \frac{kT}{m}$, si ha

$$h(v) = \sqrt{\frac{2}{\pi}}\left(\frac{m}{kT}\right)^{3/2} e^{\frac{-v^2 m}{2kT}} v^2.$$

La velocità più probabile (moda) si trova come il punto di massimo della funzione di distribuzione: $v_p = \sqrt{\frac{2kT}{m}}$. Inoltre, la velocità media sarà

$$\bar{v} = \langle v\rangle = E[v] = \int_0^\infty v h(v)dv = \sqrt{\frac{8kT}{\pi m}}.$$

$$E[v^2] = \frac{3\pi}{8}\bar{v}^2.$$

N.d.r.: L'ultimo risultato è particolarmente interessante, visto che $E[v^2]$ è proporzionale all'energia cinetica media, che risulta perciò collegata al quadrato della velocità media.

2.12 Più variabili casuali

Se abbiamo variabili casuali $x_1, \ldots, x_n$, la funzione di distribuzione congiunta è $f(x_1, \ldots, x_n)$. Si ha

$$dP(x_1^* < x_1 < x_1^* + dx_1, \ldots, x_n^* < x_n < x_n^* + dx_n) = f(x_1^*, \ldots, x_n^*)dx_1 \cdots dx_n$$

Se $A_i = (x_i^* - \Delta x_i, x_i^* + \Delta x_i)$,

$$P(x_1 \in A_i, \ldots, x_n \in A_n) = \int_{A_1} dx_1 \cdots \int_{A_n} dx_n\, f(x_1, \ldots, x_n)$$

La condizione di normalizzazione diventa: $\int_{\Omega_1} dx_1 \cdots \int_{\Omega_n} dx_n\, f(x_1, \ldots, x_n) = 1$.

In generale la congiunta è difficile da trattare, ma se le variabili sono indipendenti, $P(A \wedge B) = P(A) \cdot P(B)$, per cui

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n)$$

Ad esempio, se le variabili sono indipendenti e con distribuzione normale $N(\mu_i, \sigma_i^2)$, si ha

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x_i-\mu_i)^2/2\sigma_i^2} = \frac{1}{(2\pi)^{n/2}\sigma_1 \cdots \sigma_n} e^{-\sum_i (x_i-\mu_i)^2/2\sigma_i^2}$$

La funzione di distribuzione marginale per la variabile x_i è la proiezione sull'asse x_i della congiunta:

$$f_{Mi}(x_i) = \int_{\Omega_1} dx_1 \cdots \int_{\Omega_{i-1}} dx_{i-1} \int_{\Omega_{i+1}} dx_{i+1} \cdots \int_{\Omega_n} dx_n \, f(x_1, \ldots, x_n).$$

La funzione di distribuzione condizionata (fissati valori su tutte le altre variabili) è

$$f_{c_i}(x_i) = c f(x_1^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_n^*) = \frac{f(x_1^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_n^*)}{\int_{\Omega_i} dx_i f(x_1^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_n^*)}.$$

Il valore di aspettazione e la varianza rispetto ad una variabile sono

$$E[x_i] = \mu_i = \int_{\Omega_1} \int_{\ldots} \int_{\Omega_n} dx_1 ... dx_n x_i f(x_1, ..., x_n) = \int_{\Omega_1} dx_1 \cdots \int_{\Omega_n} dx_n \, x_i f(x_1, \ldots, x_n)$$

$$\mathrm{var}(x_i) = \sigma_i^2 = \int \cdots \int_{\Omega} dx_1 \ldots dx_n f(x_1, \ldots, x_n)(x_i - \mu_i)^2 = E[x_i^2] - \mu_i^2$$

Si definisce inoltre

$$\mathrm{cov}(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)] = E[x_i x_j] - E[x_i]E[x_j] =$$

Coefficiente di correlazione:

$$\rho_{ij} = \frac{\mathrm{cov}(x_i, x_j)}{\sigma_i \sigma_j}$$

Nota: $-1 \leq \rho \leq 1$. Questa disuguaglianza è dimostrabile in più modi, tra i quali osservando che $E[xy]$ è un prodotto scalare $\langle x, y \rangle$, perché è bilineare, simmetrico e definito positivo. Quindi $\rho_{ij} = \frac{\langle y_i, y_k \rangle}{\|y_i\| \|y_j\|} = \cos\theta \in [-1, 1]$.

La correlazione è detta:

- Debole per $\rho_{ij} \in [-0.3, 0.3]$
- Media per $\rho_{ij} \in [-0.7, -0.3] \cup [0.3, 0.7]$
- Forte per $\rho_{ij} \in [-1, -0.7] \cup [0.7, 1]$

2.13 Matrice delle covarianze

La matrice delle covarianze V sintetizza i concetti di covarianza e di varianza. Infatti, i suoi elementi sono proprio le covarianze e quelli diagonali sono le varianze:

$$V_{ij} = \rho_{ij}\sigma_i\sigma_j$$

$$V = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \ldots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \rho_{n2}\sigma_n\sigma_2 & \cdots & \sigma_n^2 \end{pmatrix}$$

dove $\rho_{ij} = \rho_{ji}$.

2.14 Esempi di distribuzioni congiunte

Ora vediamo due esempi, uno relativo a variabili casuali discrete e un altro con quelle continue.

- Distribuzione multinomiale (variabili discrete): è una generalizzazione della distribuzione binomiale (cfr 2.3). Se invece ci sono più di due esiti possibili, la distribuzione binomiale non è sufficiente e bisogna usare quella multinomiale, costruita in modo molto simile.

  Supponiamo che l'evento possa verificarsi con n modalità (esiti) e che la probabilità che si verifichi con la k-esima modalità sia p_k. La somma delle probabilità deve essere pari a 1. Quindi, supponiamo che si verifichino k_1 eventi secondo la modalità 1, k_2 eventi secondo la 2 e così via. La somma di questi k_i deve essere pari a N. Quindi, si considera una possibile sequenza di esiti che verifichi queste ipotesi. Per esempio:

  $$1, 1, 1, ..., 1, 2, 2, ..., 2, 3, 3, ...$$

  dove 1 (prima modalità) si ripete per k_1 volte, 2 (seconda modalità) per k_2 volte e così via. Se gli eventi sono indipendenti, la probabilità di ottenere ciascun evento secondo una certa modalità è:

  $$p_i^{k_i}$$

  Se si combinano più modalità, si deve fare il prodotto delle singole probabilità. Quindi, la probabilità di ottenere un'intera sequenza è:

  $$p(k_1, \ldots, k_n) = \frac{N!}{k_1! \ldots k_n!} p_1^{k_1} \cdots p_n^{k_m}.$$

  che corrisponde alla distribuzione multinomiale.

  Il valore di aspettazione è:

  $$E[k_i] = Np_i,$$

  mentre la varianza è:

  $$\text{var}(k_i) = Np_i(1 - p_i)$$

  Ha senso che siano espressioni molto simili a quelli della distribuzione binomiale perché quest'ultima è un caso particolare di quella multinomiale. Inoltre, la covarianza vale:

  $$\text{cov}(k_i, k_j) = Np_ip_j,$$

  che risulta essere diversa da zero. Un valore maggiore di p_i comporta un valore minore di p_j e viceversa. In aggiunta, più le due probabilità sono piccole, più tende a zero la covarianza e, di conseguenza, più sono indipendenti le due variabili casuali. Infine, il coefficiente di correlazione è:

  $$\rho_{ij} = -\frac{Np_ip_j}{N\sqrt{p_i(1-p_i)p_j(1-p_j)}} = -\left(\frac{p_ip_j}{(1-p_i)(1-p_j)}\right)^{1/2}$$

  Alcuni esempi:

  1. Supponiamo che $p_i = 0.9$ e $p_j = 0.1$. Allora:

     $$\rho_{ij} = -\left(\frac{0.9 \cdot 0.1}{0.1 \cdot 0.9}\right)^{\frac{1}{2}} = -1$$

     ovvero c'è una correlazione negativa completa, il che è ragionevole perché abbiamo solo due esiti possibili in cui la somma delle rispettive probabilità fa proprio 1.

2. Supponiamo che $p_i = 0.1$ e $p_j = 0.1$. Allora:

$$\rho_{ij} = -\frac{0.1}{0.9} \approx -0.11$$

La distribuzione multinomiale descrive perfettamente il numero di eventi all'interno di ciascun intervallo di un istogramma, perché tiene conto della possibile correlazione. Infatti, ricordiamo che le distribuzioni binomiali e di Poisson comunque rappresentano un'approssimazione. Dunque, se non ho più di metà degli eventi in un unico intervallo, posso usare questa distribuzione perché ho abbastanza varietà.

Ora consideriamo un intervallo di tempo Δt, sia N il numero totale di conteggi nell'intervallo:

$$N = \sum_{j=1}^{m} k_j$$

sia ν il suo valore di aspettazione. La probabilità di avere N eventi all'interno dell'intervallo di tempo è, quindi:

$$P(N) = \frac{\nu^N e^{-\nu}}{N!}$$

Invece, la probabilità di avere N eventi che si distribuiscono secondo le n modalità è pari al prodotto tra la distribuzione di Poisson e quella multinomiale:

$$P(N; k_1, ..., k_n) = \frac{\nu^N \cdot e^{-\nu}}{k_1! \cdot \ldots k_n!} p_1^{k_1} \ldots p_n^{k_n}$$

che posso riscrivere come:

$$P(N; k_1, ..., k_n) = \frac{e^{-\nu(p_1+\cdots+p_n)}}{k_1! \cdot \cdots \cdot k_n!} \cdot (\nu p_1)^{k_1} \ldots (\nu p_n)^{k_n} = \frac{e^{-\nu p_1}(\nu p_1)^{k_1}}{k_1!} \cdot \frac{e^{-\nu p_2}(\nu p_2)^{k_2}}{k_2!} \ldots \frac{e^{-\nu p_n}(\nu p_n)^{k_n}}{k_n!}$$

e quindi posso scriverla come prodotto di n distribuzioni di Poisson:

dove $\nu_i = \nu \cdot p_i$.

- Distribuzione multinormale (variabili continue con funzione di distribuzione gaussiana):

  Siano $x_1, x_2, ..., x_n$ variabili casuali che seguono tutte una distribuzione gaussiana definita da:

  $$f_i(x_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x_i-\mu_i)^2/2\sigma_i^2}$$

  Se sono indipendenti, la funzione di distribuzione congiunta è fattorizzata, ovvero è il prodotto delle funzioni di distribuzioni relative alle singole variabili.

  $$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x_i-\mu_i)^2/2\sigma_i^2} = \frac{1}{(2\pi)^{n/2}\sigma_1 \cdots \sigma_n} e^{-\sum_i (x_i-\mu_i)^2/2\sigma_i^2}$$

  Si può usare la matrice delle covarianze per scriverla in un altro modo:

  $$V = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

Il cui determinante è:

$$\det(V) = \sigma_1^2 \cdot \sigma_2^2 \dots \sigma_n^2 = \prod_{i=1}^{n} \sigma_i^2$$

mentre la matrice inversa è:

$$V^{-1} = \begin{pmatrix} \sigma_1^{-2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^{-2} \end{pmatrix}$$

Ora applico tale matrice al vettore definito come differenza tra le variabili casuali e i rispettivi valori di aspettazione:

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad \vec{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

Quindi:

$$V^{-1}(\vec{x} - \vec{\mu}) = V^{-1} \begin{pmatrix} x_1 - \mu_1 \\ \vdots \\ x_n - \mu_n \end{pmatrix} = \begin{pmatrix} \frac{x_1 - \mu_1}{\sigma_1^2} \\ \vdots \\ \frac{x_n - \mu_n}{\sigma_n^2} \end{pmatrix}$$

Da cui si ricava la forma quadratica:

$$Q^2 = (\vec{x} - \vec{\mu})^T \cdot V^{-1}(\vec{x} - \vec{\mu}) = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma_i^2}$$

che ci serve per scrivere in modo compatto la funzione di distribuzione multinormale:

$$f(x_1, ..., x_n) = \frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{(\det V)^{1/2}} e^{-Q^2/2}$$

La distribuzione ha valore massimo quando $Q^2 = 0$ e ha dei valori costanti quando Q^2 assume dei valori costanti. Consideriamo il caso bidimensionale:

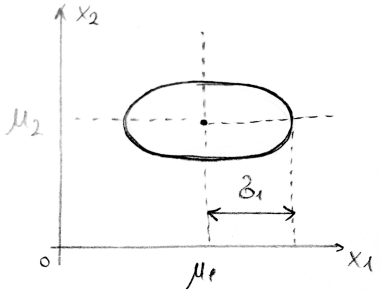


Figura 1: Ellisse: $Q^2 = 1$; centro: Q^2=0

Da cui si possono ricavare un po' di probabilità:

$$P(\mu_1 - \sigma_1 \leq x_1 \leq \mu_1 + \sigma_1) = 0.68$$

$$P(\mu_1 - 2\sigma_1 \leq x_1 \leq \mu_1 + 2\sigma_1) = 0.95$$

$$P(nel\ rettangolo\ di\ 1\ \sigma) = 0.68^2$$

$$P(nell'ellisse) = P(Q^2 \leq 1) = P(\chi_2^2 \leq 1) = \int_0^1 \frac{1}{2} e^{-z/2} \approx 0.39$$

Se, invece, considero $Q^2 = 4$ ottengo l'ellisse che come semiassi ha proprio il doppio delle deviazioni standard. La probabilità di ottenere un valore al suo interno è:

$$P(Q^2 \leq 4) = P(\chi_2^2 \leq 4) \approx 0.86$$

Se, invece, fossi in n dimensioni, l'ellisse diventerebbe un ellissoide $n - dimensionale$ e la distribuzione sarebbe χ_n^2.

Tutte queste considerazioni si possono fare anche quando le variabili considerate non sono indipendenti. Nel caso generale, la matrice delle covarianze non è più diagonale; restringendoci al caso $n = 2$ il suo determinante è:

$$det(V) = \sigma_1^2 \sigma_2^2 (1 - \rho^2)$$

La sua inversa è:

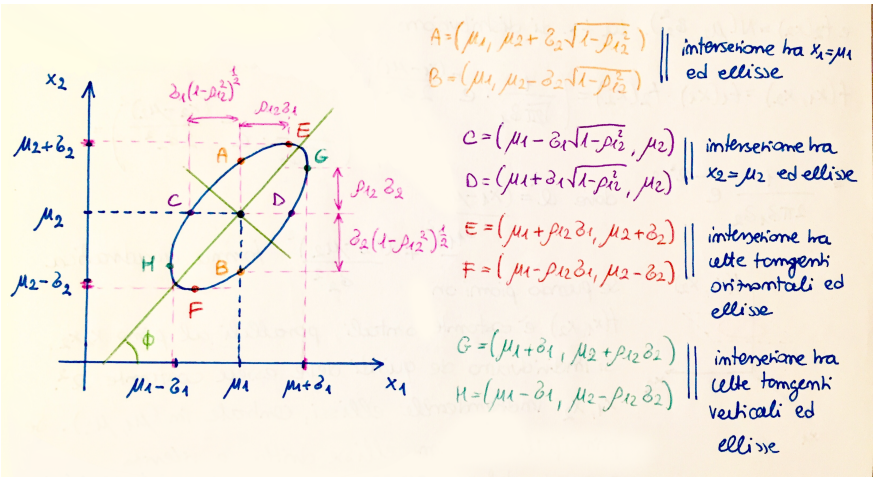
$$V^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1 \sigma_2} \\ \frac{-\rho}{\sigma_1 \sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}$$

Come prima, la applico a $\vec{x} - \vec{\mu}$:

Da cui ottengo il termine Q^2:

$$Q^2 = \frac{1}{1-\rho^2} \left(\frac{(x_1 - \mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1 \sigma_2} + \frac{(x_2 - \mu_2)}{\sigma_2^2} \right)$$

Se $Q^2 = 1$: Se considero la funzione condizionata di x_2, si trova che la funzione di distri-



buzione di x_2 è ancora gaussiana che però che ha valore massimo che si muove sulla retta di regressione (CD). Il valore di aspettazione risulta essere:

$$\mu_2' = \rho \frac{\sigma_2}{\sigma_1} (x_1 * -\mu_1) + \mu_2$$

Mentre la varianza:

$$\sigma_2'^2 = \sigma_2^2(1-\rho^2)$$

Quindi, se $\rho = 0$, la varianza rimane la stessa, altrimenti risulta essere più piccola rispetto a quella riferita alla distribuzione marginale.

Nonostante Q^2 abbia questa espressione, continua ad avere una distribuzione χ^2_2. Inoltre la probabilità nel rettangolo è sempre la stessa (perché si ricava dalle due marginali, non tiene conto della correlazione)

2.15 Funzioni generatrici dei momenti

Analogamente al caso di una variabile, definiamo le funzioni generatrici dei momenti per una funzione di distribuzione di n variabili casuali:

La funzione generatrice dei momenti algebrici:

$$M^*(t_1,\dots,t_n) = E\left[e^{\sum_i t_i x_i}\right] = \int_{\Omega_1} dx_1 \cdots \int_{\Omega_n} dx_n\, e^{\sum_i t_i x_i} f(x_1,\dots,x_n)$$

La funzione generatrice dei momenti centrali:

$$M(t_1,\dots,t_n) = E\left[e^{\sum_i t_i (x_i-\mu_i)}\right] = \int_{\Omega_1} dx_1 \cdots \int_{\Omega_n} dx_n\, e^{\sum_i t_i (x_i-\mu_i)} f(x_1,\dots,x_n)$$

In questo caso però si può anche ricavare la covarianza tra due variabili casuali:

$$\mathrm{cov}(x_i,x_j) = \left.\frac{\partial^2 M(t_1,\dots,t_n)}{\partial t_i \partial t_j}\right|_{\vec{t}=0}$$

3 Funzioni di variabili casuali: valore di aspettazione e varianza

Sia $x_1, \ldots, x_n$ un campione di variabili casuali. Esso può essere costituito da una serie di misure ripetute di una grandezza fisica X, caratterizzata da una funzione di distribuzione $f(\lambda_1, \ldots, \lambda_m)$ che dipende da un certo numero di parametri $\lambda_1, \ldots, \lambda_m$. Se si potessero ripetere le misure infinite volte, ovvero se si potessero avere a disposizione infiniti campioni, allora si potrebbe determinare la f e i suoi parametri senza alcun dubbio. Tuttavia, nella pratica si ha sempre un campione limitato dal quale, tramite l'inferenza statistica, si riescono a ottenere informazioni sull'intera popolazione. L'obiettivo principale è ottenere delle stime dei parametri che, necessariamente, devono essere funzioni del campione dal momento che sono ricavate proprio da quest'ultimo.

Dunque, è evidente l'importanza delle funzioni di più variabili casuali per la stima dei parametri, ma rivestono un ruolo altrettanto rilevante anche all'interno del test di ipotesi. In entrambi i casi si costruiscono delle funzioni del campione, chiamate "funzioni statistiche", che ci permettono di ottenere informazioni sull'intera popolazione.

3.1 Caso di una variabile

Sia x una variabile casuale con valore di aspettazione μ_x e varianza σ_x^2. Sia y una variabile che ci interessa e che sia una funzione di x, ovvero $y = y(x)$. Se si conosce la funzione di distribuzione $f(x)$ della variabile x, si potrebbe essere interessati a conoscere anche quella di y. Tuttavia, a volte ciò non è necessario ed è sufficiente conoscere il suo valore di aspettazione μ_y e la sua varianza σ_y^2. Questi ultimi si possono ottenere attraverso uno sviluppo in serie in un intorno di μ_x:

$$y(x) \cong y(\mu_x) + \frac{dy}{dx}\Big|_{x=\mu_x}(x-\mu_x) + \frac{1}{2}\frac{d^2y}{dx^2}\Big|_{x=\mu_x}(x-\mu_x)^2 + \ldots$$

Ci fermiamo al secondo ordine: dopo ci sono termini di ordine superiore in $x - \mu_x$. Calcoliamo il valore di aspettazione dell'intera espressione:

$$\mu_y \cong y(\mu_x) + \frac{1}{2}\frac{d^2y}{dx^2}\Big|_{x=\mu_x}\sigma_x^2$$

Quindi, così si ottiene il valore di aspettazione di y.

Per quanto riguarda la varianza di y:

$$\mathrm{var}(y) = \sigma_y^2 = E[y^2] - (E[y])^2 = E[y^2] - \mu_y^2$$

Per riscrivere il primo addendo, dobbiamo ottenere un'espressione approssimata per y^2, attraverso lo sviluppo in serie scritto prima:

$$y^2 \cong y^2(\mu_x) + \left(\frac{dy}{dx}\Big|_{x=\mu_x}\right)^2(x-\mu_x)^2 + 2y(\mu_x)\frac{dy}{dx}\Big|_{x=\mu_x}(x-\mu_x) + \cancel{2}y(\mu_x)\frac{1}{\cancel{2}}\frac{d^2y}{dx^2}\Big|_{x=\mu_x}(x-\mu_x)^2$$

dove non sono stati scritti i termini superiori al secondo. Come prima, si calcola il valore di aspettazione dell'intera espressione:

$$E[y^2] \cong y^2(\mu_x) + \left(\frac{dy}{dx}\Big|_{x=\mu_x}\right)^2\sigma_x^2 + y(\mu_x)\frac{d^2y}{dx^2}\Big|_{x=\mu_x}\sigma_x^2$$

Ritornando al calcolo di σ_y^2:

$$\begin{aligned}
\sigma_y^2 &= E[y^2] - \mu_y^2 \approx \\
&\approx y^2(\mu_x) + \left(\frac{dy}{dx}\Big|_{x=\mu_x}\right)^2 \sigma_x^2 + y(\mu_x)\frac{d^2y}{dx^2}\Big|_{x=\mu_x}\sigma_x^2 - \left(y(\mu_x) + \frac{1}{2}\frac{d^2y}{dx^2}\Big|_{x=\mu_x}\sigma_x^2\right)^2 \approx \\
&\approx \cancel{y^2(\mu_x)} + \left(\frac{dy}{dx}\Big|_{x=\mu_x}\right)^2 \sigma_x^2 + \cancel{y(\mu_x)\frac{d^2y}{dx^2}\Big|_{x=\mu_x}\sigma_x^2} - \cancel{y^2(\mu_x)} - \cancel{y(\mu_x)\frac{d^2y}{dx^2}\Big|_{x=\mu_x}\sigma_x^2} = \\
&= \left(\frac{d^2y}{dx^2}\Big|_{x=\mu_x}\right)^2 \sigma_x^2
\end{aligned}$$

che è la relazione nota come legge di propagazione della varianza per una variabile casuale.

3.2 Caso di più variabili

Si procede in modo simile ma sviluppando in serie di potenze rispetto a ciascuna variabile casuale. Siano queste ultime $x_1, \ldots, x_n$ con valori di aspettazione μ_i e varianze σ_i^2. Sia y una funzione delle nostre n variabili casuali:

$$y = y(x_1, \ldots, x_n)$$

Per semplicità, introduciamo i vettori:

$$\vec{x} = (x_1, \ldots, x_n), \qquad \vec{\mu} = (\mu_1, \ldots, \mu_n)$$

Quindi, lo sviluppo in serie diventa:

$$y(\vec{x}) \cong y(\vec{\mu}) + \sum_i \frac{\partial y}{\partial x_i}\Big|_{\vec{x}=\vec{\mu}}(x_i - \mu_i) + \frac{1}{2}\sum_i\sum_j \frac{\partial^2 y}{\partial x_i \partial x_j}\Big|_{\vec{x}=\vec{\mu}}(x_i - \mu_i)(x_j - \mu_j)$$

Quindi, si calcola il valore di aspettazione dell'intera espressione:

$$E[y] = \mu_y \cong y(\vec{\mu}) + \frac{1}{2}\sum_i\sum_j \frac{\partial^2 y}{\partial x_i \partial x_j}\Big|_{\vec{x}=\vec{\mu}} \mathrm{cov}(x_i, x_j) \tag{1}$$

Nel caso di variabili indipendenti, per $i \neq j$ le covarianze sono tutte nulle mentre per $i = j$ diventano le varianze. Dunque, l'espressione del valore di aspettazione diventa:

$$\mu_y \approx y(\vec{\mu}) + \frac{1}{2}\sum_i \frac{\partial^2 y}{\partial x_i^2}\Big|_{\vec{x}-\vec{\mu}}\sigma_i^2$$

che si riconduce facilmente al caso di una variabile casuale visto nella sezione precedente.

Per quanto riguarda la varianza:

$$\mathrm{var}(y) = \sigma_y^2 = E[y^2] - (E[y])^2$$

dove come prima bisogna riscrivere il primo addendo attraverso uno sviluppo in serie di potenze di y^2. I calcoli sono i medesimi, anche se più complicati. Mettendo tutto assieme si trova:

$$\sigma_y^2 \cong \sum_i \left(\frac{\partial y}{\partial x_i}\Big|_{\vec{x}=\vec{\mu}}\right)^2 \sigma_i^2 + \sum_{i\neq j} \frac{\partial y}{\partial x_i}\Big|_{\vec{x}=\vec{\mu}} \frac{\partial y}{\partial x_j}\Big|_{\vec{x}=\vec{\mu}} \rho_{ij}\sigma_i\sigma_j \tag{2}$$

Da notare che in quest'ultima relazione il secondo termine (assente per variabili indipendenti) potrebbe essere sia positivo che negativo, al contrario del primo che è sempre positivo. Quindi, tale termine aggiuntivo potrebbe aumentare o diminuire l'errore statistico rispetto a quello calcolato trascurando il contributo delle covarianze.

3.3 Esempi

Esempio 1: somma di due variabili casuali.

Siano x_1 e x_2 due variabili casuali aventi valori di aspettazione μ_1 e μ_2, varianze σ_1^2 e σ_2^2 e covarianza $\text{cov}(x_1, x_2) = \rho\sigma_1\sigma_2$. Si definisce la variabile casuale y come $y = x_1 + x_2$. Si ottiene il suo valore di aspettazione applicando la (1):

$$E[y] = \mu_y = \mu_1 + \mu_2$$

perché le derivate parziali seconde sono nulle. Per quanto riguarda la varianza, si applica la (2):

$$\text{var}(y) = \sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2$$

Compare un 2 davanti al coefficiente di correlazione perché bisogna considerare sia la coppia 1-2 che la coppia 2-1.

Se $\rho > 0$, allora la varianza della somma è maggiore della semplice somma delle varianze. In questo caso, si ha una correlazione positiva tra x_1 e x_2, quindi, se una è maggiore del valore di aspettazione corrisponde, ciò si riflette anche nella seconda. Viceversa, se $\rho < 0$, allora la varianza della somma è minore della somma delle varianze e le due variabili sono inversamente correlate.

Esempio 2: differenza di due variabili casuali.

Analogamente al precedente esempio, definendo $y = x_1 - x_2$, si trova che:

$$\mu_y = \mu_1 - \mu_2, \qquad \sigma_y^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$$

Se $\rho > 0$, la varianza della differenza è minore della differenza delle varianze. Viceversa, se $\rho < 0$, la varianza della differenza è maggiore della differenza delle varianze.

Questo esempio è molto utile quando si vuole valutare la bontà della stima dei parametri. Queste ultime vengono ottenute a partire dallo stesso campione tramite due diversi metodi. Supponiamo che i risultati siano $\hat{m}_1$ e $\hat{m}_2$ e ci chiediamo se siano tra loro compatibili valutandone la loro differenza:

$$\delta = \hat{m}_1 - \hat{m}_2$$

e confrontandola con la loro incertezza statistica. Quindi:

$$\sigma_\delta^2 = \sigma_{\hat{m}_1}^2 + \sigma_{\hat{m}_2}^2 - 2\rho\sigma_{\hat{m}_1}\sigma_{\hat{m}_2}$$

È ragionevole supporre che il coefficiente di correlazione sia circa 1 dal momento che i due metodi di stima dei parametri dovrebbero essere altrettanto validi e, quindi, dovrei ottenere lo stessi valore. Dunque:

$$\sigma_\delta^2 \approx \sigma_{\hat{m}_1}^2 + \sigma_{\hat{m}_2}^2 - 2\sigma_{\hat{m}_1}\sigma_{\hat{m}_2} = (\sigma_{\hat{m}_1} - \sigma_{\hat{m}_2})^2$$

O equivalentemente:

$$\sigma_\delta = |\sigma_{\hat{m}_1} - \sigma_{\hat{m}_2}|$$

che, in genere, è un numero molto piccolo.

Esempio 3: media aritmetica.

Siano $x_1, \ldots, x_n$ le variabili casuali che costituiscono il campione, $\mu_1, \ldots, \mu_n$ i loro valori di aspettazione, $\sigma_1^2, \ldots, \sigma_n^2$ le loro varianze. La media aritmetica è definita come:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Se supponiamo che le x_i siano indipendenti, vale che:

$$E[\bar{x}] = \mu_{\bar{x}} = \frac{1}{n}\sum_i \mu_i, \qquad \text{var}(\bar{x}) = \sigma^2_{\bar{x}} = \frac{1}{n^2}\sum_i \sigma_i^2$$

Se i valori di aspettazione e le varianze sono tutte uguali, allora:

$$\mu_{\bar{x}} = \mu, \qquad \sigma^2_{\bar{x}} = \frac{1}{n}\sigma^2$$

Dal momento che è interessante conoscere la distanza tra le singole variabili e la loro media, si definiscono le variabili y_i nel modo seguente:

$$y_i = x_i - \bar{x}$$

Vogliamo valutare l'errore statistico sulle y_i in modo da confrontarlo con le y_i stesse. Dunque:

$$E[y_i] = E[x_i] - E[\bar{x}] = \mu - \mu = 0$$

$$\text{var}(y_i) = \text{var}(x_i - \bar{x}) = \sigma_i^2 + \frac{1}{n}\sum_i \sigma_i^2 - 2\,\text{cov}(x_i, \bar{x})$$

dove si è usata l'espressione della varianza trovata per la differenza tra due variabili casuali. La covarianza si calcola come:

$$\text{cov}(x_i, \bar{x}) = E[x_i\bar{x}] - E[x_i]E[\bar{x}] = E[x_i\bar{x}] - \mu^2$$

Il primo addendo può essere scritto come:

$$E[x_i\bar{x}] = \frac{1}{n}\sum_j E[x_i x_j] = \frac{1}{n}\sum_j (\text{cov}(x_i, x_j) + E[x_i]E[x_j]) = \frac{1}{n}((\sigma_i^2 + 0) + \sum_j \mu^2) = \frac{\sigma_i^2}{n} + \mu^2$$

Dunque la covarianza diventa:

$$\text{cov}(x_i, \bar{x}) = \frac{\sigma_i^2}{n} + \mu^2 - \mu^2 = \frac{\sigma_i^2}{n}$$

E quindi la varianza:

$$\text{var}(y_i) = \sigma_i^2 + \frac{1}{n^2}\sum_i \sigma_i^2 - 2\frac{\sigma_i^2}{n} = \sigma_i^2 + \sigma^2_{\bar{x}} - 2\frac{\sigma_i^2}{n} = \sigma_i^2\left(1 - \frac{2}{n}\right) + \sigma^2_{\bar{x}}$$

Se $n \gg 1$, il termine $2/n$ è trascurabile e la varianza sulla media è molto minore delle varianze sulle singole variabili. In simboli:

$$\sigma^2_{\bar{x}} \ll \sigma_i^2$$

Dunque:

$$\text{var}(y_i) \approx \sigma_i^2$$

Se $n = 2$, allora si ha la media di due soli elementi e la varianza sulle y_i vale:

$$\text{var}(y_i) = \sigma^2_{\bar{x}}$$

Per esempio, se y_1 è definito come:

$$y_1 = x_1 - \frac{1}{2}(x_1 + x_2) = \frac{x_1 + x_2}{2}$$

allora la varianza risulta:

$$\sigma^2_{y_1} = \frac{\sigma_1^2 + \sigma_2^2}{4} = \sigma^2_{\bar{x}}$$

Per tutti gli altri casi, bisogna tenere conto del termine $2/n$ che diminuisce il contributo delle σ_i^2. Se le x_i, in particolare, seguono una distribuzione gaussiana, la seguono anche le y_i e il rapporto y_i/σ_{y_i} segue una distribuzione normale standard. Controlli di questo tipo sono alla base della valutazione della bontà dei metodi usati all'interno dell'inferenza statistica.

Esempio 4: due funzioni delle stesse variabili casuali.

Siano $x_1, \ldots, x_n$ n variabili casuali indipendenti, $g(x_1, \ldots, x_n)$ e $h(x_1, \ldots, x_n)$ funzioni diverse però delle stesse variabili. Si trovano facilmente il valore di aspettazione e la varianza associata espressi tramite quelli delle variabili. Più interessante è la covarianza. Infatti, dal momento che le due funzioni si basano sulle stesse variabili, ci si aspetta che siano in qualche modo correlate.

$$\mathrm{cov}(g,h) = E[gh] - E[g]E[h] = \cdots = \sum_i \frac{\partial g}{\partial x_i}\Big|_{\vec{x}=\vec{\mu}} \frac{\partial h}{\partial x_i}\Big|_{\vec{x}=\vec{\mu}} \sigma_i^2 + \sum_{i \neq j} \frac{\partial g}{\partial x_i}\Big|_{\vec{x}=\vec{\mu}} \frac{\partial h}{\partial x_j}\Big|_{\vec{x}=\vec{\mu}} \mathrm{cov}(x_i, x_j)$$

dove i calcoli sono stati omessi ma si basano sempre su sviluppi in serie di potenze delle funzioni g e h, semplificando i termini oltre il secondo ordine.

3.4 Il valore atteso del prodotto come prodotto scalare

$$\begin{aligned}
\mathrm{cov}(g,h) = \langle g - E[g], h - E[h] \rangle = \langle g, h \rangle - \langle g, E[h] \rangle - \langle E[g], h \rangle + \langle E[g], E[h] \rangle = \\
= E[gh] - E[gE[h]] - E[E[g]h] + E[E[g]E[h]] = \\
= E[gh] - E[g]E[h] - E[g]E[h] + E[g]E[h] = \\
= E[gh] - E[g]E[h].
\end{aligned}$$

4 Funzioni di variabili casuali: funzione di distribuzione

4.1 Caso di una variabile

Valore di aspettazione e varianza non descrivono completamente una funzione di distribuzione, a volte non sono sufficienti per il problema che si sta affrontando.

Supponiamo di avere due variabili casuali x e y, una funzione dell'altra ($y = y(x)$), e di voler calcolare proprio la funzione di distribuzione di quest'ultima.

Allora, in questo caso, la probabilità infinitesima si scrive come:

$$\begin{aligned}
dP = P(y \in (y^*, y^* + dy)) = g(y^*)dy = \\
= P(x \in (x^*, x^* + dx)) = f(x^*)dx
\end{aligned}$$

Da cui si può ottenere la relazione tra le due funzioni di distribuzione (se la corrispondenza tra le due variabili è biunivoca):

$$g(y) = f(x) \left| \frac{dx}{dy} \right| \tag{3}$$

Se la corrispondenza non è biunivoca, si inverte y localmente e si sommano le derivate delle varie funzioni $x_k = x_k(y)$. Questa relazione l'abbiamo usata più volte senza saperlo. Per esempio, se una variabile x segue una distribuzione normale $N(\mu_x, \sigma_x^2)$ e se ne definisce un'altra come:

$$y = \frac{x - \mu_k}{\sigma_x}$$

allora quest'ultima segue una distribuzione normale standard. Oppure, sia x come prima, definiamo $y = \frac{1}{x}$. Qual è la funzione di distribuzione di y? Applichiamo la (3):

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

Esercizio utile da fare: vedere le funzioni di distribuzione di $\frac{1}{\hat{q}}$ e di $\frac{1}{\hat{m}}$ e confrontare con gli istogrammi contenenti i reciproci delle stime $\hat{q}$ e $\hat{m}$.

Altro esempio interessante: supponiamo che x segua una distribuzione normale standard (anche se non fosse così, sappiamo come ricondurci a tale distribuzione) e definiamo un'altra variabile casuale come il suo quadrato, cioè $y = x^2$. Graficamente abbiamo una parabola con concavità verso l'alto e vertice centrato nell'origine del sistema di assi. In tal caso, abbiamo due possibili inversioni: $x = \pm\sqrt{y}$, con $|\frac{dx}{dy}| = \frac{1}{2}y^{-\frac{1}{2}}$.

Applicando la (3), si trova che la funzione di distribuzione di y è la somma delle due funzioni (uguali) di y:

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

che è la funzione di distribuzione di χ^2 a un grado di libertà.

4.2 Caso di più variabili

Abbiamo visto il caso di una funzione di una sola variabile casuale. Ora complichiamoci la vita e consideriamo una funzione di più variabili casuali. In questo caso non possiamo applicare la (3) ma dobbiamo introdurre n funzioni delle n variabili casuali, che devono essere invertibili rispetto alle x_i.

Dalla funzione di distribuzione congiunta delle variabili $x_1, \ldots, x_n$ posso passare a quella delle variabili $y_1, \ldots, y_n$. Essa è definita, in modo analogo al caso unidimensionale, come:

$$g(y_1, \ldots, y_n) = f(x_1, \ldots, x_n)|J|, \tag{4}$$

dove J è lo jacobiano della trasformazione. Scrivendolo esplicitamente:

$$g(y_1, \ldots, y_n) = f(x_1, \ldots, x_n) \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix}$$

Per ottenere la funzione di distribuzione marginale, cioè quella legata alla variabile che ci interessa, è sufficiente integrare in tutte le variabili tranne proprio quella che ci interessa. In simboli:

$$g_1(y_1) = \int dy_2 \ldots dy_n g(y_1 \ldots y_n)$$

4.3 Esempi

Questa procedura non sempre è semplice e può essere evitata usando le funzioni generatrici dei momenti. Facciamo alcuni esempi importanti:

Esempio 1 Supponiamo di avere due variabili casuali indipendenti x_1 e x_2 aventi funzione di distribuzione uniforme tra 0 e 1:

$$\Omega_{x1} = \Omega_{x2} = [0, 1]$$

$$x_i \in \Omega_i$$

$$f_i(x_i) = 1$$

Poiché le due variabili sono indipendenti, la funzione di distribuzione congiunta, definita su $\Omega_{x1} \times \Omega_{x2}$, è fattorizzabile:

$$f(x_1, x_2) = f(x_1)f(x_2) = 1$$

cioè vale 1 nel quadrato di lato 1 e vertice in basso a sinistra posto nell'origine degli assi. Ora definiamo una nuova variabile come $y = x_1 + x_2$ e vogliamo calcolare la sua funzione di distribuzione. Devo introdurre y_2 funzione sempre di x_1 e x_2 ma diversa da y_1. Per esempio, $y_2 = x_1 - x_2$. Esplicito x_1 e x_2 dall'espressione di y_1 e da quella di y_2:

$$\begin{cases} x_1 = \frac{1}{2}(y_1 + y_2) \\ x_2 = \frac{1}{2}(y_1 - y_2) \end{cases}$$

Calcolo lo jacobiano della trasformazione: $|J| = \frac{1}{2}$ Quindi, usando la (4), si ottiene:

$$g(y_1, y_2) = \frac{1}{2}$$

$$g_1(y_1) = \int_{\Omega_{y2}} g(y_1, y_2) dy_2$$

Bisogna definire il dominio di tale funzione, cioè il corrispondente del quadrato di prima nel piano $y_1 - y_2$. Il punto (0,0) rimane (0,0); il punto (1,0) diventa (1,1); il punto (0,1) diventa (1,-1); infine il punto (1,1) diventa (2,0). Quindi:

$$y_2 \in \begin{cases} [-y_1, y_1] & y_1 \in [0, 1] \\ [-y_1 + 2, y_1 + 2] & y_1 \in [1, 2] \end{cases}$$

integrando devo distinguere i due casi:

$$g_1(y_1) = \begin{cases} y_1 & y_1 \in [0, 1] \\ 2 - y_1 & y_1 \in [1, 2] \end{cases}$$

cioè la funzione di distribuzione è quella triangolare. La funzione di distribuzione della differenza segue lo stesso andamento e lo si può verificare facilmente.

Se considero un grande numero di variabili, ottengo un qualcosa che si avvicina sempre di più alla funzione di distribuzione di Gauss.

Esempio 2 Date le due variabili indipendenti x_1, x_2 con distribuzione uniforme in $[0, 1]$, possiamo verificare che la distribuzione delle variabili definite come

$$y_1 = \sqrt{-2\ln(x_1)}\cos(2\pi x_2)$$

$$y_2 = \sqrt{-2\ln(x_1)}\sin(2\pi x_2)$$

(ovvero le trasformate di Box-Mueller) sia normale standard e che esse siano indipendenti tra loro.

Esplicitiamo x_2:

$$\frac{y_2}{y_1} = \tan(2\pi x_2) \implies x_2 = \frac{1}{2\pi} \arctan\left(\frac{y_2}{y_1}\right)$$

Per x_1:

$$y_1^2 + y_2^2 = -2\ln x_1 \implies x_1 = e^{-\frac{1}{2}(y_1^2+y_2^2)}$$

A questo punto, calcolo le derivate:

$$\frac{\partial x_1}{\partial y_1} = e^{-\frac{1}{2}(y_1^2+y_2^2)}(-y_1)$$

$$\frac{\partial x_1}{\partial y_2} = e^{-\frac{1}{2}(y_1^2+y_2^2)}(-y_2)$$

$$\frac{\partial x_2}{\partial y_1} = \frac{1}{2\pi} \cdot \frac{1}{1+\frac{y_2^2}{y_1^2}} \left(-\frac{y_2}{y_1}\right)$$

$$\frac{\partial x_2}{\partial y_2} = \frac{1}{2\pi} \cdot \frac{1}{1+\frac{y_2^2}{y_1^2}} \cdot \frac{1}{y_1}$$

E passiamo al determinante:

$$\frac{1}{2\pi} e^{-\frac{1}{2}(y_1^2+y_2^2)} \left[-\frac{y_1^2}{y_1^2+y_2^2} - \frac{y_2^2}{y_1^2+y_2^2}\right] = -\frac{1}{2\pi} e^{-\frac{1}{2}(y_1^2+y_2^2)}$$

E quindi la funzione di distribuzione congiunta è:

$$g(y_1, y_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(y_1^2+y_2^2)} = \frac{1}{\sqrt{2\pi}} e^{-y_1^2/2} \frac{1}{\sqrt{2\pi}} e^{-y_2^2/2}$$

Ovvero il prodotto di due funzioni di distribuzione normali standard; essendo fattorizzabile è verificata l'indipendenza. Quindi, in alternativa alla generazione di numeri casuali che abbiamo fatto per esercizio, se ne possono generare due e calcolare come abbiamo visto y_1 e y_2.

Esempio 3 Osserviamo ora la distribuzione data dal rapporto tra due variabili u e v indipendenti e con distribuzione $N(0,1)$. Alla nuova variabile $x_1 = \frac{u}{v}$ associamo una seconda $x_2 = v$ (arbitraria), da cui si ricava $|J| = x_2$

$$f(u,v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} = \frac{1}{2\pi} e^{-x_2^2 \frac{(1+x_1^2)}{2}}$$

Per cui

$$g(x_1, x_2) = \frac{1}{2\pi} e^{-x_2^2 \frac{(1+x_1^2)}{2}} |x_2|$$

Ottengo $h(x_1)$ come distribuzione marginale $g_{M1}(x_1)$:

$$h(x_1) = \int_{-\infty}^{+\infty} g(x_1, x_2) dx_2 = -\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-x_2^2 \frac{(1+x_1^2)}{2}} x_2 dx_2 + \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x_2^2 \frac{(1+x_1^2)}{2}} x_2 dx_2$$

$$= \frac{1}{2\pi}\left(\frac{1}{1+x_1^2}\right) e^{-x_2^2\frac{(1+x_1^2)}{2}}\Big|_{-\infty}^{0} - \frac{1}{2\pi}\left(\frac{1}{1+x_1^2}\right) e^{-x_2^2\frac{(1+x_1^2)}{2}}\Big|_{0}^{+\infty} = \frac{1}{\pi}\left(\frac{1}{1+x_1^2}\right)$$

Ovvero la funzione di distribuzione di Cauchy:

$$h(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$

Per essa $E[x]$ e la sua varianza non sono definiti.

Ponendo $x = \frac{y-y_0}{\Gamma}, y = \Gamma x + y_0$ si ottiene:

$$g(y) = \frac{1}{\pi}\frac{1}{1+\frac{(y-y_0)^2}{\Gamma^2}}\frac{1}{\Gamma} = \frac{1}{\pi}\frac{\Gamma}{\Gamma^2+(y-y_0)^2} = \frac{1}{\pi}\frac{\Gamma}{\Gamma^2+(y-y_0)^2}$$

anche nota come distribuzione di Breit-Wigner. Come detto prima, invece di usare (2) spesso risulta più semplice dimostrare che le funzioni generatrici dei momenti coincidono(Teorema sui momenti).

4.4 Somma di Variabili casuali

Consideriamo ora $y = \sum_i^n x_i$, dove le x_i sono indipendenti. Per la fattorizzabilità della loro pdf $f(x_1, \ldots, x_n)$:

$$M_y^*(t) = \prod_i^n \int_{\Omega_i} e^{tx_i} f_i(x_i) dx_i = \prod_i^n M_{x_i}^*(t)$$

- Nel caso le x_i abbiano distribuzione $\chi^2_{\nu_i}$, dove $M^*_{x_i}(t) = (1-2t)^{-\frac{\nu_i}{2}}$, risulta

$$M_y^*(t) = (1-2t)^{-\frac{\nu}{2}} \qquad \text{dove } \nu = \sum_i \nu_i$$

- Segue che: prese delle x_i con distribuzione $N(\mu_i, \sigma_i^2)$, se consideriamo $y = \sum_i \frac{(x_i-\mu_i)^2}{\sigma_i^2}$, la sua distribuzione risulta essere χ^2_n (perché abbiamo dimostrato che le variabili $z_i = \frac{(x_i-\mu_i)^2}{\sigma_i^2}$ seguono χ^2_1).

4.5 Media di variabili casuali

Legge dei grandi numeri. Date n variabili $x_1, \ldots, x_n$, con $E[x_i] = \mu$, $\text{var}(x_i) = \sigma^2$ finita, allora $\bar{x} = \frac{1}{n}\sum_i^n x_i$ converge in probabilità a μ, ovvero:

$$\forall \varepsilon > 0, \ \lim_n P(|\bar{x} - \mu| < \varepsilon) = 1$$

Dimostrazione. Per la varianza vale la disuguaglianza di Čebyšëv:

$$P\left(|\bar{x} - \mu| \geq k\sigma_{\bar{x}}\right) \leq \frac{1}{k^2} \implies P\left(|\bar{x} - \mu| \geq k\frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{k^2}$$

Definendo $\varepsilon = \frac{k\sigma}{\sqrt{n}}$

$$P(|\bar{x} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2 n} \implies \lim_n P(|\bar{x} - \mu| \geq \varepsilon) = 0$$

□

Teorema del limite centrale. Date n variabili $x_1, \ldots, x_n$ indipendenti, con $E[x_i] = \mu$, $\text{var}(x_i) = \sigma^2$ finita, allora per $n \to +\infty$ $\bar{x} = \frac{1}{n}\sum_i^n x_i$ ha distribuzione $N(\mu, \frac{\sigma^2}{n})$, qualunque sia la pdf degli x_i.

Dimostrazione. Dimostreremo il seguente fatto analogo: $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ tende ad avere distribuzione N(0,1) per $n \to \infty$. Partiamo dallo scrivere z come somma di variabili casuali indipendenti:

$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \sum_i^n \frac{(x_i-\mu)}{\sigma\sqrt{n}} = \sum_i^n z_i$$

$$z_i = \frac{x_i - \mu}{\sigma\sqrt{n}}$$

Osserviamo che, siccome le x_i hanno valore di aspettazione μ, vale che

$$\forall i,\ E[z_i] = 0,\ \sigma_{zi}^2 = E[z_i^2] = \frac{1}{n\sigma^2}\sigma^2 = 1/n$$

Per quanto visto prima, grazie all'indipendenza delle variabili z_i: $M_z^* = \prod_i^n M_{zi}^* = \prod_i^n E[e^{z_i t}]$. Sviluppiamo in serie l'esponenziale:

$$e^{z_i t} = 1 + z_i t + z_i^2 t^2/2 + z_i^3 t^3/6 + ...$$

Quindi,

$$\forall i,\ E[e^{z_i t}] = E[1 + z_i t + z_i^2 t^2/2 + z_i^3 t^3/6 + ...] = 1 + 0 + E[z_i^2]\frac{t^2}{2} + E[z_i^3]\frac{t^3}{6} + ... =$$

$$= 1 + \frac{t^2}{2n} + \frac{\mu_3 t^3}{\sigma^3 6n\sqrt{n}} + ...$$

Per n sufficientemente grande, siamo autorizzati a trascurare i termini successivi a $\frac{t^2}{2n}$. Allora si ottiene che $M_z^* \approx (1 + \frac{t^2}{2n})^n \xrightarrow[n\to\infty]{} e^{\frac{t^2}{2}}$ □

Nota: esiste un teorema del limite centrale ancora più generale (non lo useremo mai) che dice: $x_1, ..., x_n$ indipendenti, seguono $f_1(x_1), ..., f_n(x_n)$ con $E[x_i] = \mu_i$ e $var[x_i] = \sigma_i^2$ finite. Allora: $\bar{x}$ per $n \to +\infty$ tende a seguire la distribuzione $N(\frac{\sum_i \mu_i}{n}, \frac{\sum_i \sigma_i^2}{n})$

4.6 Somma degli scarti quadratici

Le proprietà asintotiche della media aritmetica (convergenza asintotica, teorema del limite centrale) mostrano come essa sia un buono stimatore del valore atteso.

Dato un insieme di variabili casuali $x_1, \ldots, x_n$ indipendenti che seguono una distribuzione normale con stessi valori di aspettazione e varianza μ e σ. Supponiamo di conoscere il primo. Definiamo la variabile

$$z = \sum_{i=1}^n (x_i - \mu)^2$$

allora a meno di un fattore di scala z segue una distribuzione χ_n^2. Infatti:

$$\frac{z}{\sigma^2} = \sum_{i=1}^n \frac{(x_i-\mu)^2}{\sigma^2}$$

Da cui posso ottenere che:

$$E\left[\frac{z}{\sigma^2}\right] = n \implies E[z] = n\sigma^2; \qquad \text{var}\left(\frac{z}{\sigma^2}\right) = 2n \implies var(z) = 2n\sigma^4$$

Se definisco la nuova variabile casuale $S'^2 = \frac{z}{n} = \frac{1}{n}\sum_i (x_i - \mu)^2$, allora il suo valore di aspettazione e la sua varianza saranno, rispettivamente:

$$E[S'^2] = \sigma^2 \qquad \text{var}(S'^2) = \text{var}(z)\frac{1}{n^2} = \frac{2n(\sigma^4)}{n^2} = \frac{2\sigma^4}{n}$$

Quindi la si può usare per stimare la varianza di una distribuzione a partire da un certo campione. Dal momento che la variabile S'^2 segue una distribuzione di χ^2_n e non una gaussiana, il contenuto probabilistico legato alla sua deviazione standard non è più il 68% (tende ad esserlo solo per $n \to \infty$). Ora supponiamo di non conoscere il valore di aspettazione di ciascuna variabile casuale x_i. Allora possiamo usare la media aritmetica dal momento che è un buono stimatore. Quindi:

$$\hat{\mu} = \bar{x}$$

Ora definiamo la variabile casuale $z = \sum_{i=1}^{n}(x_i - \bar{x})^2$ e vediamo come questa segua una distribuzione χ^2_{n-1} a meno di un fattore di scala. Il suo valore di aspettazione è:

$$E[z] = (n-1)\sigma^2$$

Se definisco $\frac{z}{\sigma^2}$, tale variabile segue una distribuzione χ^2_n. Ora, per introdurre uno stimatore che abbia come valore di aspettazione proprio σ^2, allora esso deve essere definito così:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Dato un certo numero di variabili casuali indipendenti x_i che seguono distribuzione normale $N(\mu, \sigma^2)$, se introduciamo altre variabili

$$t_i = \frac{x_i - \bar{x}}{\sigma_i}$$

E definendo la somma dei quadrati:

$$t = \sum_{i=1}^{n} t_i^2$$

Allora quest'ultima variabile segue una distribuzione di χ^2_{n-1}.

Dimostrazione. A partire dalle n variabili x_i posso passare a $n-1$ variabili y_i indipendenti. Per farlo, si usa una trasformazione ortogonale:

$$x_i \mapsto y_i = \sum_{j=1}^{n} a_{ij}x_j$$

dove i coefficienti della combinazione lineare devono soddisfare la richiesta:

$$\sum_{k=1}^{n} a_{ik}a_{jk} = \sum_{k} a_{ki}a_{kj} = \delta_{ij}$$

Questo implica che:

$$\sum_{k=1}^{n} a_{ik}^2 = 1$$

Se ora consideriamo:

$$\sum_{i=1}^{n} y_i^2 = \sum_i x_i^2$$

Definisco la y_1 in modo tale da tenere in considerazione la relazione che intercorre tra le x_i e $\bar{x}$:

$$y_1 = a \sum_i x_i = \frac{1}{\sqrt{n}} \sum_i x_i = \sqrt{n}\bar{x}$$

Per ottenere il valore di a:

$$a_{1i} = a, \qquad \sum_i a_{1i}^2 = 1 = na^2 \to a = \sqrt{\frac{1}{n}}$$

Per quanto riguarda gli altri y_i, quindi per $i \geq 2$:
la sommatoria di tutti i coefficienti è nulla:[2]

$$\sum_j a_{1j} = 0$$

Quindi, i valori di aspettazione delle y_i per $i \geq 2$ sono tutti altrettanto nulli.
Invece, le varianze sulle y_i si ricavano dalla loro definizione:

$$\sigma_{y_i}^2 = \sum_{j=1}^{n} a_{1j}^2 \sigma^2 = \sigma^2$$

A partire dalla definizione, si può calcolare anche la covarianza:

$$\mathrm{cov}(y_i, y_j) = E[y_i y_j] - E[y_i]E[y_j] = 0$$

Ora possiamo tornare alla variabile $t = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2}$. Scriviamola più esplicitamente:

$$t = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2} = \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} \bar{x}^2 - 2\sum_{i=1}^{n} x_i \bar{x}\right\} = \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n} y_i^2 + n\bar{x}^2 - 2n\bar{x}^2\right\} =$$

$$\frac{1}{\sigma^2}\left\{\sum_{i=1}^{n} y_i^2 - y_1^2\right\} = \frac{1}{\sigma^2}\sum_{i=2}^{n} y_i^2 = \sum_{i=2}^{n} \left(\frac{y_i}{\sigma}\right)^2.$$

che effettivamente segue una distribuzione di χ^2_{n-1}. □

[2] Questa relazione può essere dimostrata considerando la seguente sommatoria:

$$\sum_{k=1}^{n} a_{1k} a_{ik} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} a_{ik} = \delta_{1i} = 0.$$

In generale: quando partiamo da n variabili casuali indipendenti gaussiane $x_1, ...x_n$ e, per definire n nuove variabili casuali $t_1, ...t_n$, introduciamo r relazioni (quelle usate per definire le t_i, per fare un esempio. Per esempio la media aritmetica ma possono anche essere molto diverse), la somma dei quadrati passa da una distribuzione di χ^2_n a χ^2_{n-r}. Se per esempio usiamo le x_i per stimare un parametro, stiamo introducendo una relazione tra le variabili casuali e quindi il numero di gradi di libertà della distribuzione viene ridotto.

Un caso molto diverso è quando le variabili casuali x_i non sono ridondanti (cioè, non esistono delle relazioni che permettono di toglierne una, come invece fatto prima), però $cov(x_i, x_j) \neq 0$. In tal caso, il numero di gradi di libertà rimane lo stesso. Per dimostrarlo bisogna trovare una trasformazione ortogonale che permette di diagonalizzare la matrice delle covarianze. Così facendo si passa da n variabili casuali x_i a n variabili casuali y_i. Queste ultime possiedono una covarianza nulla. A questo punto, tramite un fattore di scala, si definiscono n variabili casuali z_i in modo tale che seguano una distribuzione normale.

Consideriamo il caso bidimensionale, con due variabili casuali x_1 e x_2 che seguono una distribuzione normale rispettivamente $N(\mu_1, \sigma_1^2)$ e $N(\mu_2, \sigma_2^2)$. Supponiamo che la matrice delle covarianze sia:

$$V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Vogliamo dimostrare che:

$$Q^2 = (\vec{x} - \vec{\mu})^T V^{-1} (\vec{x} - \vec{\mu})$$

(che è l'espressione da considerare quando la covarianza è diversa da zero) ha distribuzione di χ^2_2, anche se le covarianze non sono nulle. Scriviamo esplicitamente l'espressione di Q^2 (l'abbiamo già calcolata da qualche parte):

$$Q^2 = \frac{1}{1-\rho^2}\{a_1^2 - 2\rho a_1 a_2 + a_2^2\}$$

dove $a_i = \frac{x_i - \mu_i}{\sigma_i}$ con $i = 1, 2$.

Passiamo da x_1 e x_2 alle variabili y_1 e y_2 definite così:

$$y_{1,2} = \frac{1}{\sqrt{2}}(a_1 \pm a_2)$$

Quindi, definiamo altre due variabili casuali z_1 e z_2 tramite un fattore di scala:

$$z_{1,2} = \frac{1}{\sqrt{1 \pm \rho}} y_{1,2} =$$

Ora, se dimostriamo che possiamo scrivere Q^2 come:

$$Q^2 = z_1^2 + z_2^2$$

e che z_1 e z_2 sono indipendenti e seguono una distribuzione normale standard, allora dimostriamo anche che Q^2 segue una distribuzione χ^2_2. Cominciamo a fare un po' di calcoli:

$$z_1^2 + z_2^2 = \frac{1}{2(1+\rho)}(a_1^2 + a_2^2 + 2a_1a_2) + \frac{1}{2(1-\rho)}(a_1^2 + a_2^2 - 2a_1a_2) =$$

$$= \frac{1}{2(1-\rho^2)}\left[(1-\rho)a_1^2 + (1-\rho)a_2^2 + (1-\rho)2a_1a_2 + (1+\rho)a_1^2 + (1+\rho)a_2^2 - 2(1+\rho)a_1a_2\right] =$$

$$= \frac{1}{2(1-\rho^2)}(2a_1^2 + 2a_2^2 - 2 \cdot 2a_1a_2) = \frac{1}{(1-\rho^2)}(a_1^2 + a_2^2 - 2\rho a_1 a_2) = Q^2$$

Questa prima parte è fatta.

So che la funzione di distribuzione congiunta di x_1 e x_2 è quella binormale:

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{\det V}} e^{-Q^2/2}$$

Da queste voglio passare alla funzione di distribuzione congiunta di z_1 e z_2:

$$g(z_1, z_2) = f(x_1(z_1, z_2), x_2(z_1, z_2))|J|.$$

Quindi, per prima cosa, devo scrivere x_1 e x_2 in funzione di z_1 e z_2:

$$z_1\sqrt{2}\sqrt{1+\rho} = \frac{x_1 - \mu_1}{\sigma_1} + \frac{x_2 - \mu_2}{\sigma_2}$$

$$z_2\sqrt{2}\sqrt{1-\rho} = \frac{x_1 - \mu_1}{\sigma_1} - \frac{x_2 - \mu_2}{\sigma_2}$$

Faccio la somma tra le due equazioni:

$$2\frac{x_1 - \mu_1}{\sigma_1} = z_1\sqrt{2}\sqrt{1+\rho} + z_2\sqrt{2}\sqrt{1-\rho}$$

$$2\frac{x_2 - \mu_2}{\sigma_2} = z_1\sqrt{2}\sqrt{1+\rho} - z_2\sqrt{2}\sqrt{1-\rho}$$

Da cui si ottiene:

$$x_1 = \frac{\sigma_1}{\sqrt{2}}(z_1\sqrt{1+\rho} + z_2\sqrt{1-\rho}) + \mu_1$$

$$x_2 = \frac{\sigma_2}{\sqrt{2}}(z_1\sqrt{1+\rho} + z_2\sqrt{1-\rho}) + \mu_2$$

A questo punto possiamo calcolare le derivate parziali:

$$\frac{\partial x_1}{z_1} = \frac{\sigma_1}{\sqrt{2}}\sqrt{1+\rho} \qquad \frac{\partial x_1}{z_2} = \frac{\sigma_1}{\sqrt{2}}\sqrt{1-\rho}$$

$$\frac{\partial x_2}{\partial z_1} = \frac{\sigma_2}{\sqrt{2}}\sqrt{1+\rho} \qquad \frac{\partial x_2}{\partial z_2} = -\frac{\sigma_2}{\sqrt{2}}\sqrt{1-\rho}$$

da cui posso calcolare il determinante e prenderne il valore assoluto: $|J| = \sigma_1\sigma_2\sqrt{1-\rho^2}$ Quindi la funzione di distribuzione congiunta g di z_1 e z_2 è :

$$g(z_1, z_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp(-(z_1^2+z_2^2)/2)\sigma_1\sigma_2\sqrt{1-\rho^2} = \frac{1}{\sqrt{2\pi}} \exp(-z_1^2/2)\cdot\frac{1}{\sqrt{2\pi}} \exp(-z_2^2/2)$$

Visto che la funzione di distribuzione congiunta è fattorizzata e ciascun fattore è una funzione di distribuzione normale standard, allora sia z_1 che z_2 seguono quest'ultima distribuzione e sono indipendenti. Abbiamo dimostrato ciò che volevamo. Questo fatto vale in generale, non solo nel caso bidimensionale.

A questo punto torniamo alla nostra variabile S^2:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

dove x_i sono n variabili indipendenti con distribuzione normale $N(\mu, \sigma^2)$. Allora, a meno di un fattore di scala, S^2 ha una distribuzione di χ^2_{n-1}, il suo valore di aspettazione è σ^2 e la sua varianza è $(2\sigma^2)^2/(n-1)$.

Visto che la deviazione standard è più interessante, per il suo significato probabilistico, prendo la radice quadrata e i due valori caratteristici diventano:

$$E[\sqrt{S^2}] = \sigma \qquad var(\sqrt{S^2}) = (\frac{1}{2}\frac{1}{\sqrt{S^2}})^2 \cdot var(S^2) = \frac{1}{4S^2} \cdot \frac{2(\sigma^2)^2}{n-1} = \frac{1}{2(n-1)}\sigma^2$$

Dove abbiamo sostituito S^2 con il proprio valore di aspettazione. Quindi, l'errore relativo è:

$$\frac{\sigma_{\sqrt{S^2}}}{\sigma} = \frac{1}{\sqrt{2(n-1)}}$$

- Per $n = 1$ diverge, infatti non si può fare la stima della varianza usando un solo valore.
- Per $n = 2$, diventa $1/1.4 =$ troppo.
- Per $n = 3$, diventa 0.5 (se usassimo solo tre valori per stimare la varianza, avremmo un errore relativo percentuale su tale stima del 50%, che è tanto...).
- Per $n = 10$, diventa 0.24, il che va meglio. Quindi, per stimare abbastanza bene la varianza bisogna fare un numero sufficientemente alto di misure.

5 La stima dei parametri (*parameter fitting*)

La stima dei parametri è un argomento estremamente importante all'interno dell'inferenza statistica. I parametri sono quelli che caratterizzano una distribuzione oppure la relazione tra grandezze fisiche. L'obiettivo è estrarre informazioni da un set di dati che riguardano un'intera popolazione. Per il momento ci limiteremo alle stime puntuali, cioè alle stime di un solo parametro a partire da un campione (un certo numero di misure). Tuttavia, non è sufficiente stimare un valore numerico ma bisogna anche sapervi associare un'incertezza statistica adeguata. Questo è il problema legato agli intervalli di confidenza, che permettono di determinare la probabilità che il valore vero del parametro si trovi dentro a un certo intervallo.

Ci sono diversi metodi statistici per stimare correttamente i valori dei parametri, noi ne vedremo due: il metodo dei minimi quadrati (MMQ) e quello del *Maximum Likelihood* (MML). In teoria, entrambi dovrebbero fornire una buona stima, quindi dovrebbero essere equivalenti quando applicati. È importante notare che, anche se le ipotesi statistiche alla base sono sbagliate, i due metodi forniscono sempre una soluzione (per esempio, il metodo dei minimi quadrati fornisce una stima del coefficiente angolare e dell'intercetta anche se i punti sperimentali non seguono un andamento lineare). La correzione a questo aspetto viene affrontata dal test di ipotesi, che confronta il risultato trovato da una stima con le ipotesi statistiche considerate per la stessa.

Esempio: Sia x una variabile casuale descritta da una funzione di distribuzione $f(x)$. Prendiamo un campione (*sample*) di n misure indipendenti (*size* n) $x_1, ...x_n$. Queste ultime sono a loro volte delle variabili casuali, quando non ci limitiamo a un unico esperimento. Infatti, se rifacciamo l'esperimento, otterremo un altro campione che non possiamo predire. La funzione di distribuzione congiunta del campione è:

$$h(x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i)$$

dove si suppone che le funzioni di distribuzione delle singole variabili siano le stesse. Ora immaginiamo di non conoscere la $f(x)$. Vogliamo ottenere informazioni su di essa a partire dal campione di misure (**questo è il problema centrale della statistica: come ottenere informazioni sulla popolazione a partire da un suo campione limitato**). In generale, si possono fare alcune ipotesi sulla distribuzione e sui suoi parametri, pur non conoscendoli. Il vettore:

$$\vec{\theta} = (\theta_1, \theta_2, ..., \theta_\lambda)$$

che ingloba tutti i parametri che intervengono nella funzione di distribuzione. Per semplicità, per il momento, supponiamo che tale vettore sia costituito da una sola componente. L'obiettivo è costruire delle funzioni del solo campione (statistiche) dalle quali estrarre le caratteristiche principali della funzione di distribuzione, come il valore di aspettazione, la varianza, il valore dei parametri.

$$f(x, \theta) \qquad t(x_1, x_2, ..., x_n)$$

Una statistica, cioè una funzione del singolo campione, si dice "stimatore" (*estimator*) se permette di ottenere una delle caratteristiche della funzione di distribuzione. Per ogni caratteristica è necessario uno stimatore. Nel caso del parametro θ, lo stimatore $\hat{\theta}(x_1, ..., x_n)$ mi da informazioni sul valore del parametro stesso. Essendo funzioni di variabili casuali, lo stimatore è anch'esso una variabile casuale. Di conseguenza, segue una funzione di distribuzione con certe caratteristiche.

Nota: La stima è il valore numerico che lo stimatore assume sul campione considerato. Quindi, lo stimatore è una funzione mentre la stima è lo stimatore calcolato sul campione e, dunque, è un valore numerico (attenzione alla differenza, però nella pratica si usano entrambi i termini senza alcuna distinzione).

Affinché gli stimatori siano effettivamente utili, devono rispettare alcune caratteristiche minime:

- Convergenza in probabilità al valore vero del parametro:

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0$$

ovvero all'aumentare delle misure a disposizione, cioè alla dimensione del campione, il valore stimato del parametro si avvicina sempre di più al suo valore vero. Questa è il minimo requisito per uno stimatore utile. Uno stimatore che rispetta questa condizione è detto "consistente". In genere si riferisce a un multicampione di dimensione infinita.

- Il valore di aspettazione dovrebbe essere il valore vero del parametro:

$$E[\hat{\theta}] = \int \hat{\theta} \cdot g(\hat{\theta}, \theta) d\hat{\theta}$$

dove $g(\hat{\theta}, \theta)$ è la funzione di distribuzione di $\hat{\theta}$. Poiché $\hat{\theta}$ è una funzione del campione, il valore di aspettazione può anche essere calcolato come:

$$E[\hat{\theta}] = \int_{\Omega_1} ... \int_{\Omega_n} dx_1 ... dx_n \cdot \hat{\theta}(x_1, ..., x_n) \cdot f(x_1, ...x_n)$$

dove $f(x_1, ..., x_n)$ è la fz. di distribuzione congiunta.

Poiché normalmente il valore di aspettazione non coincide con il valore vero del parametro, si introduce un *bias*, definito come:

$$b = E[\hat{\theta}] - \theta$$

Se $b = 0$, lo stimatore è "centrato" o *unbiased*. Potrebbe succedere che $b \neq 0$ ma che:

$$\lim_{n\to\infty} b = 0$$

cioè quando si ha un campione di dimensione sufficientemente grande, la differenza tra i due diventa talmente piccola da essere trascurabile. In tal caso, lo stimatore si dice "asintoticamente centrato".

Nota: i concetti di stimatore consistente e asintoticamente centrato sono molto diversi. Il primo si riferisce a una situazione limite nella quale si ha un unico campione con infinite misure; il secondo, invece, impiega infiniti campioni di dimensione, però, finita dal momento che compare il valore di aspettazione. Infatti, quest'ultimo si basa su ciò che abbiamo a disposizione, ovvero un numero finito di misure.

- La varianza dello stimatore:

$$\sigma^2_{\hat{\theta}} = E[(\hat{\theta} - E(\hat{\theta}))^2]$$

anche questo si riferisce a un campione di dimensione finita. Serve per fornire l'incertezza statistica sulla stima del parametro. Un buon stimatore deve avere una varianza piccola, cioè l'errore statistico minore possibile, legata all'efficienza del metodo di stima.

- Errore quadratico medio (*Mean Squared Error*, MSE):

$$MSE = E[(\hat{\theta} - \theta)^2]$$

La differenza con la varianza è solo che, invece di sottrarre il valore di aspettazione dello stimatore, sottraggo il valore vero del parametro. La varianza e il MSE coincidono solo se lo stimatore è centrato, perché in tal caso il valore di aspettazione coincide con il valore vero del parametro. Altrimenti, il MSE vale:

$$\text{MSE} = \sigma_{\hat{\theta}}^2 + b^2$$

cioè bias e varianza si sommano quadraticamente per formare l'errore quadratico medio.

Altri criteri per classificare gli stimatori sono la semplicità e, in particolare, com'è fatta la funzione di distribuzione dello stimatore. Molto spesso si assume che quest'ultima sia gaussiana.

5.1 Metodo del *Maximum Likelihood* / della massima verosimiglianza

Sia x una variabile casuale con funzione di distribuzione $f(x, \theta)$, sia $x_1, ..., x_n$ il campione a disposizione. Supponiamo che abbiano tutte la stessa funzione di distribuzione. Quella congiunta sarà, se le variabili sono indipendenti:

$$f(x_1, \ldots, f_n) = \prod_{i=1}^{n} f(x_i, \theta) dx_i$$

ed è la probabilità che ciascuna variabile casuale stia all'interno di un determinato intervallo. Se conosciamo θ (ne fisso il valore) e la sua distribuzione, allora possiamo calcolare la produttoria e trovare, in teoria, una probabilità alta. In altre parole, visto che abbiamo trovato proprio quei valori, la probabilità di ottenerli evidentemente non può essere bassa, altrimenti saremmo stati molto fortunati per tutte quante le misure costituenti il campione. Quindi, se individuiamo correttamente la funzione di distribuzione delle misure e il valore del parametro, troviamo una probabilità alta e viceversa. Dunque, questo mi fornisce una misura di quanto bene ci stiamo approcciando alla stima.

$$\mathcal{L}(\vec{x}, \theta) = \prod_{i=1}^{n} f(x_i, \theta) dx_i$$

Il valore migliore del parametro è quello che massimizza questa funzione, ovvero che massimizza la probabilità di ottenere le misure che compaiono nel campione. Questo è il principio cardine del metodo del *Maximum Likelihood*. In genere, si considera il logaritmo della funzione di *Likelihood*, che è la stessa cosa nella pratica (diventa più semplice perché i prodotti diventano delle somme).

$$\frac{\partial \ln(\mathcal{L})}{\partial \theta} = 0$$

è la condizione da imporre per massimizzare la funzione, in cui ovviamente quest'ultima deve essere derivabile. Se siamo fortunati, si riesce a trovare la funzione $\hat{\theta}(x_1, ..., x_n)$ che annulla la derivata prima della funzione di *Likelihood*. Qualora, invece, ciò non fosse possibile, si ricorre a dei metodi numerici.

Se ci sono più parametri da stimare, si massimizza rispetto a ciascun parametro, ovvero si calcolano tutte le derivate parziali necessarie:

$$\frac{\partial \ln(\mathcal{L})}{\partial \theta_1} = 0 \quad ... \quad \frac{\partial \ln(\mathcal{L})}{\partial \theta_\lambda} = 0$$

dove tutte queste condizioni sono messe a sistema e sono chiamate "equazioni di *Likelihood*". Gli stimatori ottenuti con questo metodo soddisfano le condizioni richieste per un buon stimatore, in particolari sono centrati.

Esempio 1: variabili gaussiane

Sia x la variabile casuale a disposizione, $N(\mu, \sigma^2)$ la sua funzione di distribuzione, $x_1, ..., x_n$ il campione associato di variabili casuali indipendenti. Consideriamo la funzione di *Likelihood*, ovvero quella congiunta:

$$\mathcal{L}(x_1, ..., x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\left(-\sum_i \frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

dove μ e σ sono i parametri che vogliamo stimare. Prendiamo il logaritmo naturale dell'espressione:

$$\ln \mathcal{L} = \ln\left(\frac{1}{(2\pi)^{n/2}}\right) - \frac{n}{2}\ln(\sigma^2) - \sum_i \frac{(x_i - \mu)^2}{2\sigma^2}$$

Quindi, bisogna considerare le derivate parziali rispetto a μ e a σ^2 in modo da trovare le loro migliori stime.

$$\begin{aligned}
\frac{\partial}{\partial a}(\ln\ \mathcal{L}) &= \frac{\partial}{\partial a}\left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - a)^2\right) = \frac{\partial}{\partial a}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - a)^2\right) = \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial}{\partial a}(x_i - a)^2 = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}2(x_i - a) = 0 \Leftrightarrow \sum_{i=1}^{n}2(x_i - \hat{a}) = 0 \Leftrightarrow \\
&\Leftrightarrow \sum_{i=1}^{n}(x_i - \hat{a}) = 0 \Leftrightarrow \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\hat{a} = 0 \Leftrightarrow \sum_{i=1}^{n}x_i - n\hat{a} = 0 \Leftrightarrow \hat{a} = \frac{1}{n}\sum_{i=1}^{n}x_i
\end{aligned}$$

ovvero il migliore stimatore (centrato) del valore atteso è la media aritmetica. Passiamo alla varianza:

$$\begin{aligned}
\frac{\partial}{\partial \sigma^2}(\ln\ \mathcal{L}) &= \frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - a)^2\right) = \\
&= -\frac{n}{\cancel{2}}\frac{1}{\cancel{2}\pi\sigma^{\cancel{2}}} \cdot \cancel{2} \cdot \cancel{2\pi}\sigma + \frac{\cancel{2}\sigma}{\sigma^{\cancel{4}}} \cdot \frac{1}{\cancel{2}}\sum_{i=1}^{n}(x_i - a)^2 = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - a)^2 = 0 \Leftrightarrow \\
&\Leftrightarrow -\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma^3}}\sum_{i=1}^{n}(x_i - a)^2 = 0 \Leftrightarrow \frac{-n\hat{\sigma^2} + \sum_{i=1}^{n}(x_i - a)^2}{\hat{\sigma^3}} = 0 \\
&\Leftrightarrow -n\hat{\sigma^2} + \sum_{i=1}^{n}(x_i - a)^2 = 0 \Leftrightarrow \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2
\end{aligned}$$

ovvero la migliore stima della varianza è la media aritmetica delle varianze riferite alle singole misure. Da notare che questo stimatore, però, non è centrato. Infatti:

$$E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$$

Quindi il bias vale:

$$b = E[\hat{\sigma}^2] - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

Asintoticamente è centrato ma comunque in questa forma può dare fastidio. Per avere uno stimatore sempre centrato, lo si pone uguale:

$$S^2 = \frac{1}{n-1}\sum_i (x_i - \hat{\mu})^2$$

in quanto il suo valore di aspettazione è proprio uguale alla varianza.

Esempio 2: variabili esponenziali

Sia t la variabile casuale in questione con funzione di distribuzione esponenziale:

$$f(t,\tau) = \frac{1}{\tau} e^{-t/\tau}$$

e sia $t_1, ..., t_n$ il campione a disposizione di variabili indipendenti. Sappiamo che il valore di aspettazione è τ e la varianza è τ^2. Usiamo il MML per vedere se le stime coincidono con questi valori.

$$\mathcal{L} = \frac{1}{\tau^n} \exp\left(-\sum_{i=1}^{n} \frac{t_i}{\tau}\right)$$

Prendo il logaritmo:

$$\ln \mathcal{L} = -n \ln \tau - \sum_{i=1}^{n} \frac{t_i}{\tau}$$

Prendo la derivata parziale e la pongo pari a zero:

$$\frac{\partial \ln \mathcal{L}}{\partial \tau} = -\frac{n}{\tau} + \frac{1}{\tau^2} \sum_i t_i = 0$$

da cui si ottiene:

$$\hat{\tau} = \frac{1}{n} \sum_i t_i$$

cioè nuovamente il migliore stimatore per il valore di aspettazione è la media aritmetica. Notiamo anche che il suo valore di aspettazione coincide con il valore vero del parametro:

$$E[\hat{\tau}] = \tau$$

Nota: Nella pratica, con τ si indica la vita media, che è la media dei tempi di sopravvivenza nel caso dei decadimenti radioattivi, e si usa anche il suo reciproco $\lambda = 1/\tau$ chiamata "costante di decadimento". Molto spesso, si stima molto di più quest'ultimo. Se, invece, mi interessasse la funzione del parametro $a(\theta)$ e non tanto il parametro stesso:

$$\frac{\partial \ln L}{\partial \theta} = \frac{\partial ln L}{\partial a} \cdot \frac{\partial a}{\partial \theta} = 0$$

A parte il caso in cui la funzione a è costante, che non è molto interessante, si ottiene che la stima migliore di a è:

$$\hat{a}(\theta) = a(\hat{\theta}),$$

cioè basta valutare la funzione nella stima migliore del parametro (proprietà di invarianza per trasformazioni del *Likelihood*). Tornando ai tempi di attesa, si trova di conseguenza:

$$\hat{\lambda} = \frac{1}{\hat{\tau}} = \frac{n}{\sum_i t_i}$$

che non è un risultato intuitivo. Il suo valore di aspettazione è:

$$E[\hat{\lambda}] = \int_{\Omega_1} ... \int_{\Omega_2} \lambda \cdot \prod_i f(x_i, \lambda) dx_i = \frac{n}{n-1} \lambda,$$

ovvero, il valore non è centrato, c'è un bias non nullo. Quindi partendo da uno stimatore centrato $\hat{\tau}$ si è passati a uno non centrato semplicemente facendo il reciproco.

— **Fine esempio 2**

Come calcoliamo le incertezze statistiche legate alle stime? È utile avere un'idea della larghezza delle distribuzioni riferite ai valori puntuali delle stime effettuate a partire da diversi campioni di misure. Un modo per quantificare l'incertezza statistico è quello di considerare la varianza:

$$\sigma^2_{\hat{\theta}} = E[(\hat{\theta} - E[\hat{\theta}])^2]$$

Ci sono dei casi semplici in cui questa espressione è facile da calcolare, come per S^2 definito prima e per la media aritmetica $\bar{x}$:

$$\bar{x} \qquad \rightarrow \qquad \sigma^2_{\bar{x}} = \frac{\sigma^2_x}{n}$$

$$S^2 \qquad \rightarrow \qquad \sigma^2_{S^2} = \frac{\tau^2}{n}$$

Tuttavia, il problema è che le varianze generalmente sono espresse in funzione del parametro che vogliamo stimare, quindi di cui non conosciamo il valore a priori. Però sappiamo che vale la proprietà di invarianza per trasformazioni del *Likelihood*:

$$\hat{a}(\theta) = a(\hat{\theta})$$

E dunque:

$$x \qquad \rightarrow \qquad \hat{\sigma}^2_{\hat{\mu}} = \frac{\hat{\sigma}^2}{n}$$

$$S^2 \qquad \rightarrow \qquad \hat{\sigma}^2_{\hat{\tau}} = \frac{\hat{\tau}^2}{n}$$

cioè valuto le funzioni delle incertezze nel valore stimato del parametro. Se queste varianze sono difficili da calcolare, si usano le simulazioni di Monte-Carlo, in cui si simulano molte volte le nostre n misure come se si avessero a disposizione molti campioni, quindi si stimano i parametri per ciascuno di essi, si studia la loro distribuzione e si determina la larghezza a una deviazione standard della distribuzione "reale" a partire dalla larghezza di quella "simulata", fornendo l'intervallo che corrisponde a una probabilità del 68%). Altrimenti, si può usare la disuguaglianza di Cramer-Rao-Frechet.

Quest'ultima dice che, dato un campione, ci possono essere diversi stimatori ma la loro varianza non può essere inferiore a un certo valore, cioè c'è un limite all'incertezza statistica che si può associare a uno strumento.

Sia θ la grandezza che vogliamo stimare, $\hat{t}$ il suo stimatore e il valore di aspettazione di quest'ultimo sia funzione di θ, ovvero $E[\hat{t}] = \tau(\theta)$. Inoltre supponiamo che lo stimatore non sia centrato e, quindi, che abbia un bias pari a b. La diseguaglianza di Cramer-Rao-Frechet dice che:

$$\mathrm{var}(\hat{t}) \geq \frac{\left(\frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta}\right)^2}{E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right]}$$

e questa disuguaglianza vale sempre, qualunque sia il metodo usato per ottenere lo stimatore. Il caso più semplice è quello in cui $\tau(\theta) = \theta$ e $b = 0$. In tal caso, la disuguaglianza diventa:

$$\mathrm{var}(\hat{t}) \geq \frac{1}{E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right]}$$

Dimostrazione. Per dimostrarla, abbiamo bisogno di diverse relazioni. La prima che ci serve è la seguente:

$$\int_{\Omega_1} \cdots \int_{\Omega_2} dx_1 \dots dx_n \, \mathcal{L}(x_1, \dots, x_n; \theta) = 1$$

cioè che la funzione di *Likelihood* deve essere normalizzata. Ne consideriamo la derivata rispetto al parametro θ:

$$\frac{\partial}{\partial \theta} \int d\vec{x} \, \frac{\partial \ln \mathcal{L}}{\partial \theta} \mathcal{L} = 0$$

deve essere nulla perché l'integrale vale 1. Ma se gli estremi di integrazione non dipendono dal parametro, la relazione può essere scritta come:

$$\int_{\vec{\Omega}} d\vec{x} \, \frac{\partial \mathcal{L}}{\partial \theta} = 0$$

Inoltre, ci serve scrivere il valore di aspettazione della derivata parziale seconda rispetto al logaritmo naturale della funzione di *Likelihood* in altro modo:

$$E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right] = E\left[\left(\frac{\partial \ln \mathcal{L}}{\partial \theta}\right)^2\right]$$

Per dimostrare l'uguaglianza consideriamo:

$$E\left[\frac{\partial \ln \mathcal{L}}{\partial \theta}\right] = \int d\vec{x} \, \frac{\partial \ln \mathcal{L}}{\partial \theta} \mathcal{L} = \int d\vec{x} \, \frac{\partial \mathcal{L}}{\partial \theta} = 0$$

Derivo ora il valore di aspettazione rispetto al parametro, anche la derivata è nulla:

$$0 = \frac{\partial}{\partial \theta} \int d\vec{x} \, \frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial}{\partial \theta} \int d\vec{x} \, \frac{\partial \ln \mathcal{L}}{\partial \theta} \cdot \mathcal{L} = \int d\vec{x} \, \frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} \mathcal{L} + \int d\vec{x} \, \frac{\partial \ln \mathcal{L}}{\partial \theta} \frac{\partial \mathcal{L}}{\partial \theta} =$$
$$= \int d\vec{x} \, \frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} \mathcal{L} + \int d\vec{x} \left(\frac{\partial \ln \mathcal{L}}{\partial \theta}\right)^2 \mathcal{L} = E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] + E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]$$

da cui:

$$E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right] = E\left[\left(\frac{\partial \ln \mathcal{L}}{\partial \theta}\right)^2\right] \qquad \square$$

A quetso punto, possiamo riscrivere il termine a sinistra della disuguaglianza di Cramer-Rao-Frechet nel modo seguente:

$$var(\hat{t}) = E[(\hat{t} - \theta)^2]$$

In più, sappiamo che:

$$E[x^2]E[y^2] \geq (E[xy])^2$$

e l'uguaglianza vale solo se $x = ky$.[3] Dunque possiamo scrivere:

$$E[(\hat{t} - E[\hat{t}])^2] \cdot E\left[\left(\frac{\partial \ln \mathcal{L}}{\partial \theta}\right)^2\right] \geq \left(E\left[(\hat{t} - \theta)\frac{\partial \ln \mathcal{L}}{\partial \theta}\right]\right)^2$$

[3] L'uguaglianza vale quando x e y sono legati da una relazione lineare, come $x = ky$. In questo caso:

$$\frac{\partial \ln \mathcal{L}}{\partial \theta} = k(\hat{t} - E[\hat{t}]) \qquad \text{k costante}$$

Dobbiamo verificare che il valore di aspettazione scritto a destra faccia uno:

$$E\left[(\hat{t} - E[\hat{t}])\frac{\partial \ln \mathcal{L}}{\partial \theta}\right] = E\left[\hat{t} \cdot \frac{\partial \ln \mathcal{L}}{\partial \theta}\right] - E\left[E\left[\hat{t}\right]\frac{\partial \ln \mathcal{L}}{\partial \theta}\right]$$

Il primo addendo è:

$$E\left[\hat{t}\frac{\partial \ln \mathcal{L}}{\partial \theta}\right] = \int d\vec{x}\, \hat{t}\frac{\partial \ln \mathcal{L}}{\partial \theta} \cdot \mathcal{L} = \int d\vec{x}\, \hat{t} \cdot \frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial}{\partial \theta}\int d\vec{x}\, \hat{t} \cdot \mathcal{L} = \frac{\partial}{\partial \theta} E[\hat{t}]$$

Quando il bias è nullo (questa è la situazione che stiamo esaminando, non quella generale!), il valore di aspettazione dello stimatore è uguale al valore vero del parametro:

$$b = E[\hat{t}] - \theta = 0 \qquad \rightarrow \qquad E[\hat{t}] = \theta$$

Dunque:

$$E\left[\hat{t}\frac{\partial \ln \mathcal{L}}{\partial \theta}\right] = \frac{\partial \theta}{\partial \theta} = 1$$

L'altro addendo invece:

$$E\left[E[\hat{t}]\frac{\partial \ln \mathcal{L}}{\partial \theta}\right] = E[\hat{t}] \cdot E\left[\frac{\partial \ln \mathcal{L}}{\partial \theta}\right] = 0$$

dove l'ultimo passaggio è dovuto al fatto che per ottenere la stima del parametro si impone proprio:

$$\frac{\partial \ln \mathcal{L}}{\partial \theta} = 0$$

Quindi, si è verificata la seguente disuguaglianza:

$$E[(\hat{t} - E[\hat{t}])^2] \cdot E\left[\left(\frac{\partial \ln \mathcal{L}}{\partial \theta}\right)^2\right] \geq 1$$

che è proprio quella di Cramer-Rao-Frechet nel particolare caso considerato. □

Riassumendo, la disuguaglianza di Cramer-Rao-Frechet permette di identificare correttamente la minima incertezza statistica che può essere associata allo stimatore. Se quest'ultimo è centrale, siamo nel caso particolare dimostrato prima:

$$\sigma_{\hat{t}}^2 = \frac{1}{E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right]}$$

Il valore di aspettazione al denominatore dipende dal campione e rappresenta l'informazione del parametro contenuto nel campione, è detto "informazione di Fisher".

Ci sono tre considerazioni da fare:

1. Abbiamo un campione di n variabili casuali che seguono una funzione di distribuzione $f(x;\theta)$. Allora, la funzione di *Likelihood* è:

   $$\mathcal{L} = \prod_i f(x_i, \theta)$$

   e il suo logaritmo:

   $$\ln \mathcal{L} = \sum_i \ln(f(x_i, \theta))$$

E quindi il valore di aspettazione:

$$E\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right] = \int d\vec{x}\, \frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} \cdot \mathcal{L} = \int d\vec{x} \underbrace{\left(\frac{\partial^2}{\partial \theta^2} \sum_i \ln f(x_i, \theta)\right)}_{\text{Va come } n\frac{\partial^2}{\partial \theta^2} \ln f} \prod_i f(x_i, \theta)$$

Di conseguenza la varianza minima è inversamente proporzionale alla dimensione del campione (va come $\frac{1}{n}$). In altre parole, maggiore è la dimensione del campione e minore è la varianza minima che si può calcolare tramite la disuguaglianza di Cramer-Rao-Frechet.

2. Spesso è difficile calcolare la quantità:

$$E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right]$$

In tal caso, si sostituisce con la stessa derivata calcolata nel valore stimato del parametro:

$$E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right] \to -\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\Big|_{\theta=\hat{\theta}}$$

Ovviamente questa è un'approssimazione, non è l'espressione analitica corretta della varianza minima. Per esplicitare ciò, la varianza così calcolata si indica con:

$$\widehat{\sigma^2_{\hat{t}}}$$

3. L'espressione della varianza minima può essere generalizzata per includere anche la covarianza. Supponiamo di avere più parametri da stimare $\theta_1, \ldots, \theta_\lambda$. Allora, introducendo la matrice delle covarianze V ($cov(\hat{\theta}_i, \hat{\theta}_i) = V_{ij}$) e la sua inversa V^{-1}, si trova che:

$$(V^{-1})_{ij} = E\left[-\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln \mathcal{L}\right]$$

Esempio 1: variabili gaussiane con varianze diverse

Consideriamo un campione di n elementi $x_1, \ldots, x_n$ dove ciascuno ha valore di aspettazione $E[x_i] = \mu$ e varianza $var(x_i) = \sigma_i^2$ (con varianza nota, $\forall i$). Supponiamo che abbiano tutti distribuzione normale $N(\mu, \sigma_i^2)$. Vogliamo stimare μ a partire dal campione. Come prima cosa, dunque, scriviamo la funzione di *Likelihood*:

$$\mathcal{L} = \frac{1}{(2\pi)^{\frac{n}{2}} \cdot \sigma_1 \ldots \sigma_n} \cdot \exp\left(-\sum_i \frac{(x_i - \mu)^2}{2\sigma_i^2}\right)$$

Ne prendiamo il logaritmo naturale, di cui ci interessa solo la parte che dipende dal parametro dato che poi dobbiamo fare la derivata parziale rispetto a quest'ultimo:

$$\ln \mathcal{L} = c - \sum_i \frac{(x_i - \mu)^2}{2\sigma_i^2}$$

Ponendo uguale a zero la derivata parziale rispetto a μ troviamo:

$$\hat{\mu} = \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}$$

Ma possiamo anche scrivere:

$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = \left(\sum_i \frac{1}{\sigma_i^2} \right) (\hat{\mu} - \mu)$$

dove vediamo che si manifesta una relazione lineare tra $\frac{\partial \mathcal{L}}{\partial \mu}$ e $(\hat{\mu} - \mu)$. Quindi, possiamo calcolare la varianza minima:

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \mu^2} = -\sum_i \frac{1}{\sigma_i^2} \qquad \text{da cui} \qquad \sigma_{\hat{\mu}}^2 = \frac{1}{E\left[\sum_i \frac{1}{\sigma_i^2}\right]} = \frac{1}{\sum_i \frac{1}{\sigma_i^2}}$$

Esempio 2: variabili gaussiane con varianze uguali

È interessante anche il caso analogo in cui le varianze delle singole variabili siano uguali. In tal caso, è utile considerare sia la stima del valore di aspettazione che quella della varianza a partire dallo stesso campione. Poiché il campione è uno, le due stime risultano correlate. Ripetendo il procedimento si ottiene:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (x_i - \hat{\mu})^2$$

Si possono ricavare anche le incertezze $\sigma_{\hat{\mu}}^2$ e $\sigma_{\hat{\sigma}}^2$ sulle stime, rispettivamente, $\hat{\mu}$ e $\hat{\sigma}$ sapendo che:

$$var(\chi_\nu^2) = 2\nu$$

Infine, si può verificare che la covarianza tra $\hat{\mu}$ e $\hat{\sigma^2}$ è nulla.

Esempio 3: campione con distribuzione esponenziale

Come prima ma il campione segue una distribuzione esponenziale. In tal caso la funzione di *Likelihood* è:

$$\mathcal{L} = \frac{1}{\tau^n} \exp(-\sum_i \frac{x_i}{\tau})$$

Da cui:

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \tau^2} = \frac{n}{\tau^2} - \frac{2n\hat{\tau}}{\tau^3}$$

da cui il risultato noto:

$$\widehat{\sigma_{\hat{\tau}}^2} = \frac{\hat{\tau}^2}{n}$$

che non dovrebbe sorprendere dato che:

$$\hat{\tau} = \frac{1}{n} \sum_i x_i$$

Esempio 4: campione con distribuzione di Poisson

Come prima ma con la distribuzione di Poisson, ovvero:

$$P_k = \frac{\nu^{k_i} e^{-\nu}}{k_i!} \qquad E[k_i] = \nu \qquad var(k_i) = \nu$$

La funzione di *Likelihood* è:

$$\mathcal{L}(k_1, \ldots, k_n; \nu) = \prod_{i=1}^{n} \frac{\nu^{k_i} e^{-\nu}}{k_i!}$$

Ne calcoliamo il logaritmo:

$$\ln \mathcal{L} = \sum_{i=1}^{n} (-\ln(k_i!) + k_i \cdot \ln(\nu) - \nu)$$

Quindi, la derivata parziale rispetto a ν:

$$\frac{\partial \ln \mathcal{L}}{\partial \nu} = \left(\frac{1}{\nu} \sum_i k_i \right) - n$$

Ponendola pari a zero e risolvendola rispetto al parametro, si ottiene:

$$\hat{\nu} = \frac{1}{n} \sum_i k_i$$

E quindi posso riscrivere la derivata come:

$$\frac{\partial \ln \mathcal{L}}{\partial \nu} = \frac{n\hat{\nu}}{\nu} - n = \frac{n}{\nu}(\hat{\nu} - \nu)$$

Sussiste la relazione di linearità per la varianza minima. Dunque la derivata seconda:

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \nu^2} = -\frac{n\hat{\nu}}{\nu^2}$$

5.1.1 Metodo grafico/numerico

Raramente è facile calcolare analiticamente le stime e le incertezze a esse associate. Perciò, si ricorre a metodi numerici, per cui è utile sviluppare in serie di potenze la funzione di *Likelihood* in un intorno della stima del parametro:

$$\ln \mathcal{L}(\theta) \approx \ln \mathcal{L}(\hat{\theta}) + \frac{\partial \ln \mathcal{L}}{\partial \theta}\Big|_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2} \frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\Big|_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \dots$$

Il primo addendo lo si riscrive per definizione, il secondo è nullo perché lo calcoliamo proprio nella stima del parametro, il terzo lo si riscrive nell'ipotesi di avere la varianza minima. Inoltre, si ferma lo sviluppo al secondo ordine:

$$\ln \mathcal{L}(\theta) \approx \ln \mathcal{L}_{max} - \frac{1}{2\sigma^2_{\hat{\theta}}} (\theta - \hat{\theta})^2$$

Quindi, in prima approssimazione si trova un andamento parabolico in un intorno del valore stimato del parametro, di vertice il valore stimato del parametro e sull'asse delle ordinate il logaritmo naturale della funzione di *Likelihood*. Di conseguenza, per individuare numericamente il massimo, si può calcolare $\ln \mathcal{L}$ per diverse stime dello stesso parametro e vedere attorno a quale valore si addensano.

Se per caso ci troviamo nella situazione:

$$\ln L = \ln L_{MAX} - \frac{1}{2}$$

allora vuol dire che:

$$\frac{(\theta - \hat{\theta})^2}{\sigma_{\hat{\theta}^2}} = 1$$

e quindi:

$$(\theta - \hat{\theta})^2 = \sigma_{\hat{\theta}^2} \quad \rightarrow \quad \theta - \hat{\theta} = \sigma_{\hat{\theta}} \quad \rightarrow \quad \theta = \hat{\theta} \pm \sigma_{\hat{\theta}}$$

In altre parole è possibile determinare la deviazione standard sulla stima del parametro tracciando la retta alla quota $\ln(\mathcal{L}_{max}) - \frac{1}{2}$ e individuando i punti di intersezione, invece di usare la relazione di Cramer-Rao-Frechet.

La parabola poteva essere vista anche considerando la funzione di *Likelihood* come funzione di distribuzione del parametro sul campione:

$$\mathcal{L}(\theta) = \mathcal{L}_{max} \cdot \exp\left(-\frac{(\theta - \hat{\theta})^2}{2\sigma^2_{\hat{\theta}}}\right)$$

ed è scritta così perché normalmente i valori stimati seguono, almeno asintoticamente, una distribuzione gaussiana. La costante moltiplicativa è proprio $\mathcal{L}_{max}$ perché deve essere verificata la condizione:

$$\mathcal{L}(\hat{\theta}) = \mathcal{L}_{max}$$

Nel caso di più parametri, l'espressione diventa::

$$L(\vec{\theta}) = \mathcal{L}_{max} \cdot e^{-\frac{Q^2}{2}} \qquad \text{con} \qquad Q^2 = (\vec{\theta} - \vec{\hat{\theta}})^T \cdot V^{-1}_{\hat{\theta}} (\vec{\theta} - \vec{\hat{\theta}})$$

Quindi, nella pratica si prende un intervallo per entrambi i parametri, si fa una griglia, si calcola la fz. di Likelihood per tutti i punti individuati, si calcola quello massimo e quindi le stime dei parametri (essenzialmente quello che abbiamo fatto nell'esercitazione 5).

Se si fissa un parametro lasciando che l'altro vari (quindi, nel caso si abbiano solo due parametri da stimare), i logaritmi della funzione di *Likelihood* seguono tutti un andamento parabolico i cui vertici insieme descrivono un'altra parabola più larga.

5.1.2 *Binned Maximum Likelihood*

Vediamo ora un uso un po' diverso ma molto comune e importante del MML. Supponiamo di avere un campione $x_1, \ldots, x_n$ dove $n \gg 1$, ovvero abbiamo a disposizione un campione di dimensione molto elevata. Supponiamo inoltre che le x_i seguano tutte una stessa funzione di distribuzione $f(x_i, \theta_i)$ che dipende da un parametro θ_i. Quindi, tutte le x_i si riferiscono a un'unica grandezza fisica x con una certa funzione di distribuzione $f(x, \vec{\theta})$. L'obiettivo è stimare i parametri di quest'ultima.

Quando si ha un campione molto ampio, generalmente si istogrammano i valori e, di conseguenza, si passa da un campione di n elementi a uno di $m < n$ che coincidono con il numero di eventi (conteggi) che cadono in ciascun intervallo dell'istogramma. Per applicare il MML bisogna essere in grado descrivere la funzione di distribuzione congiunta delle nuove m variabili che costituiscono il nuovo campione. Essa sarà di tipo multinomiale:

$$\mathcal{L}(n_1, \ldots, n_m; \vec{\theta}) = \frac{n!}{n_1! \ldots n_m!} \, p_1^{n_1} \ldots p_m^{n_m}$$

dove n è il numero totale di eventi, n_i il numero di eventi che cade nell'i-esimo intervallo e p_i la probabilità associata. In simboli:

$$p_i = p\left(x_i^* - \frac{\Delta}{2} \leq x \leq x_i^* + \frac{\Delta}{2}\right) = \int_{x_i^* - \frac{\Delta}{2}}^{x_i^* + \frac{\Delta}{2}} f(x, \vec{\theta}) dx \simeq f(x_i^*, \vec{\theta})\Delta$$

dove Δ è l'ampiezza dell'intervallo e x_i^* è il suo valore centrale. Se l'ampiezza dell'intervallo è sufficientemente piccola allora si può scrivere:

$$p_i \approx f(x_i^*, \vec{\theta})\Delta$$

ovvero la probabilità dipende dai parametri che dobbiamo stimare. Da qua procediamo come sempre, calcolando il logaritmo naturale della funzione di *Likelihood*, prendendone la derivata e ponendola pari a zero, trovando così la migliore stima dei parametri.

Nota: A differenza del metodo dei minimi quadrati, con questo metodo non è un problema avere degli intervalli dell'istogramma con pochi o addirittura nessun evento. Infatti, in tal caso il contributo alla funzione di *Likelihood* è semplicemente 1 e la lascia inalterata.

5.1.3 *Extended Maximum Likelihood*

È un'altra variante del MML applicata agli istogrammi e si riferisce al caso quando il numero totale di eventi n sia considerato non fisso ma una variabile casuale, il cui valore di aspettazione è ν. Di conseguenza, il valore di aspettazione riferito al singolo intervallo dell'istogramma è:

$$E[n_i] = \nu p_i = \nu_i$$

In questo caso, la funzione di distribuzione congiunta è il prodotto di m distribuzioni di Poisson:

$$f(n_1 \dots n_m) = \prod_{i=1}^{m} \frac{\nu_i \, e^{-\nu_i}}{n_i!}$$

5.2 Metodo dei minimi quadrati

Consideriamo un campione di n variabili casuali y_i indipendenti che hanno tutte distribuzione normale $N(\mu_i, \sigma_i^2)$. Inoltre, supponiamo che i valori di aspettazione dipendano sia dai parametri che vogliamo stimare, sia da un'altra variabile x_i che viene calcolata in corrispondenza di ciascuna y_i e che supponiamo essere priva di errori (non è una variabile casuale, è solamente una variabile). Per esempio, una relazione lineare soddisfa una situazione di questo tipo:

$$E[y] = mx + q \to \mu_i = E[y_i] = \mu(x_i; \vec{\theta})$$

Se le varianze σ_i^2 sono tutte note e se le variabili y_i sono indipendenti tra di loro, la funzione di distribuzione congiunta è:

$$f(y_1, \dots, y_n) = \mathcal{L} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(y_i - \mu_i)^2}{2\sigma_i^2}}.$$

Prendendone il logaritmo naturale:

$$\ln \mathcal{L} = \ln \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \right) - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma_i^2}$$

e massimizziamolo. Per farlo, dobbiamo rendere minima il secondo addendo:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma_i^2}$$

Il metodo dei minimi quadrati dice che la stima migliore dei parametri è quella che minimizza sempre questa sommatoria, **indipendentemente dalla funzione di distribuzione delle** y_i. Ciò rappresenta un enorme vantaggio perché vuol dire che non bisogna per forza conoscere la fz. di distribuzione delle y_i per poter stimare correttamente i parametri.

Per estensione, se si prendono delle variabili con covarianza non nulla, quindi dipendenti le une dalle altre:

$$X^2 = (\vec{y} - \vec{\mu})^T \, V^{-1} \, (\vec{y} - \vec{\mu})$$

Considerazione #1: le stime ottenute con questo metodo hanno le stesse caratteristiche di quelle che si ottengono con il MML (almeno asintoticamente centrate, consistenti, varianza centrata, ...). Inoltre,i due metodi dovrebbero fornire delle stime $\hat{\theta}_L$ e $\hat{\theta}_Q$ simili, se non uguali. La loro differenza è talmente piccola che non è confrontabile con l'incertezza statistica associata alle stime.

Considerazione #2: Se le ipotesi sui valori di aspettazione delle y_i sono corrette, allora il valore minimo della sommatoria è una variabile casuale (ha un valore specifico solo su un campione specifico). Se le y_i hanno distribuzione normale $N(\mu_i, \sigma_i^2)$, e **solo in questo caso**, allora la sommatoria è una variabile di χ^2_ν con $\nu = n - k$ (n è la dimensione del campione, k è il numero di relazioni che esistono tramite il campione stesso che, in questo caso, sono le stime dei parametri). Per questo motivo, anche se quanto detto vale solo quando le y_i seguono distribuzione gaussiana, il metodo dei minimi quadrati viene chiamato anche "metodo di minimizzazione del χ^2".

Considerazione #3: se le varianze sono tutte uguali fra di loro, la forma quadratica da minimizzare diventa:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma^2}$$

dove il fattore $1/\sigma^2$ è del tutto ininfluente nel processo di minimizzazione, essendo costante. Quindi è sufficiente considerare il resto dell'espressione. Questo risulta un vantaggio nel momento in cui non si conosce σ^2 perché tanto non compare nell'equazione.

Considerazione #4: Si può assumere che le stime siano a varianza minima. Se la varianza minima non può essere calcolata, similmente a quanto fatto con il MML, per ottenere le deviazioni standard:

$$X^2 = X^2_{min} + 1 \qquad \text{1 deviazione standard}$$

$$X^2 = X^2_{min} + 4 \qquad \text{2 deviazioni standard}$$

$$X^2 = X^2_{min} + 9 \qquad \text{3 deviazioni standard}$$

— **Fine considerazione #4**

Si può verificare che le stime che sono state ottenute analiticamente con il MML sono simili a quelle che si ottengono con il MMQ.

Esempio:

In particolare, si può considerare il caso in cui le variabili casuali y_i hanno tutte lo stesso valore di aspettazione:

$$E[y_i] = \mu$$

e che le varianze σ_i^2 siano note. Se μ non è noto, potrebbe essere il parametro da stimare. La forma quadratica da minimizzare è in questo caso:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma_i^2}$$

Bisogna calcolarne la derivata e porla uguale a zero:

$$\frac{dX^2}{d\mu} = 0$$

Si ottiene la media pesata, coerentemente con quanto trovato con il MML.
— **Fine esempio**

Il metodo dei minimi quadrati può essere usato sia per misure dirette che per istogrammi. Supponiamo di avere un campione $x_1, \ldots, x_n$ di dimensione $n \gg 1$ (questa condizione equivale ad avere abbastanza dati da poterli disporre in un istogramma) e che si riferisce a una distribuzione di parametri che vogliamo stimare. In questo caso, non posso applicare il metodo ai singoli elementi del campione e devo passare a un campione di dimensione minore m, istogrammando i valori a disposizione. Con n_i denotiamo il numero di eventi nel singolo intervallo dell'istogramma $x_i^* \pm \frac{\Delta}{2}$. Il valore di aspettazione di ciascuno di questi è:

$$E[n_i] = np_i = n \int_{x_i^* - \frac{\Delta}{2}}^{x_i^* + \frac{\Delta}{2}} f(x, \vec{\theta}) dx \approx n f(x_i^*, \vec{\theta}) \Delta$$

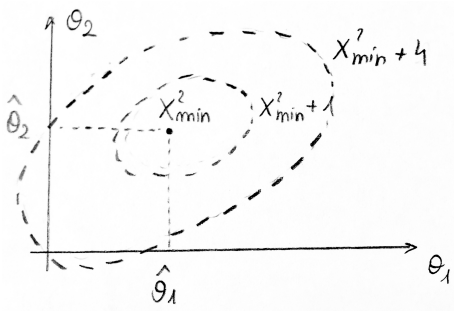
Osserviamo che tali valori di aspettazione dipendono dai parametri. Applichiamo a questo punto il metodo dei minimi quadrati e scriviamo la forma quadratica da minimizzare:

$$X^2 = \sum_{j=1}^{m} \frac{(n_j - E[n_j])^2}{E[n_j]}$$

dove è stato supposto che $p_i \ll 1$ ovvero che valga (circa) Poisson, con $\sigma_{n_i}^2 \approx np_i$. Quindi si calcolano le derivate parziali della forma quadratica rispetto a ciascun parametro e si annullano tutte quante:

$$\frac{\partial X^2}{\partial \theta_1} = 0 \qquad \ldots \qquad \frac{\partial X^2}{\partial \theta_\lambda} = 0$$

Se non si riescono a ottenere le stime $\hat{\theta}_1, \ldots, \hat{\theta}_\lambda$ dei parametri analiticamente, si procede con dei metodi numerici più o meno sofisticati con i quali si trova il minimo della forma quadratica e, quindi, gli intervalli di confidenza corrispondenti a una, due e tre deviazioni standard.



5.2.1 MMQ semplificati

Nel caso di un istogramma, la forma quadratica da minimizzare è:

$$X^2 = \sum_{j=1}^{m} \frac{(n_j - E[n_j])^2}{E[n_j]}$$

Tuttavia, spesso è scomodo avere $E[n_j]$ al denominatore perché dipende dai parametri da stimare. Di conseguenza, spesso al posto di questo valore di aspettazione si mette il numero di conteggi misurato:

$$X^2 \approx \sum_{j=1}^{m} \frac{(n_j - E[n_j])^2}{n_j}$$

Tuttavia, per effettuare questa approssimazione è necessario che il numero di eventi all'interno dell'intervallo considerato non sia troppo basso, perché altrimenti si stimerebbe male la varianza.

Nota: Al contrario del MML, qua è problematico avere degli intervalli con pochi eventi.

5.2.2 MMQ lineare

C'è un caso particolare del MMQ che permette di ottenere sempre la soluzione analitica e delle relazioni (proprietà) che sono esatte, in tal caso, e approssimate, in generale. In questo metodo i valori di aspettazione delle variabili casuali dipendono linearmente dai parametri:

$$\vec{\mu} = A\vec{\theta}$$

dove A è una matrice che non deve dipendere dai parametri. Per esempio, vanno bene relazioni come:

$$y = \theta_0 + \theta_1 x, \qquad y = \theta_0 + \theta_1 x + \theta_2 x^2$$

perché si può scrivere, rispettivamente:

$$\vec{\mu} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \vec{\theta}, \qquad \vec{\mu} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \vec{\theta}$$

La dipendenza da x può anche non essere lineare, l'importante è che lo sia quella dei parametri. Se ho a disposizione le relazioni:

$$\mu_i = E[y_i], \qquad \vec{\mu} = A\vec{\theta}$$

allora posso scrivere la forma quadratica come:

$$X^2 = (\vec{y} - A\vec{\theta})^T V^{-1} (\vec{y} - A\vec{\theta})$$

di cui devo fare la derivata rispetto a $\vec{\theta}$ e porla uguale a zero:

$$\frac{\partial X^2}{\partial \vec{\theta}} = 0$$

Ovvero, ricordando che V è una matrice simmetrica:

$$(\vec{y} - A\vec{\theta})^T V^{-1} A = 0 \leftrightarrow A^T V^{-1} (\vec{y} - A\vec{\theta}) = 0$$

Dunque:

$$A^T V^{-1} \vec{y} - A^T V^{-1} A \vec{\theta} = 0$$

da cui, applicando $(A^T V^{-1} A)^{-1}$

$$\hat{\vec{\theta}} = (A^T V^{-1} A)^{-1} A^T V^{-1} \vec{y} = B\vec{y}$$

dove $\vec{\hat{\theta}}$ è una matrice di dimensione $\lambda \times 1$, il primo fattore ha dimensione $\lambda \times n$ mentre il secondo ha dimensione $n \times 1$. Le migliori stime dei parametri sono quelle ottenute applicando alle variabili casuali che costituiscono il campione la matrice B ("del disegno") nota. La soluzione c'è sempre, a meno di matrici singolari, e quindi i parametri possono sempre essere noti a priori. L'equazione scritta esprime il fatto che le stime sono combinazioni lineari del campione. Di conseguenza, ne condividono la funzione di distribuzione.

Per questo particolare metodo, si ricavano anche le espressioni delle varianze e delle covarianze. Se si definisce come $V_{\hat{\theta}}$ la matrice delle covarianze, l'elemento generico è:

$$(V_{\hat{\theta}})_{ij} = E[(\hat{\theta}_i - E[\hat{\theta}_i])(\hat{\theta}_j - E[\hat{\theta}_j]]$$

E quindi posso scrivere la matrice come:

$$V_{\hat{\theta}} = E[(\vec{\hat{\theta}} - E[\vec{\hat{\theta}}])(\vec{\hat{\theta}} - E[\vec{\hat{\theta}}])^T] = E[(B\vec{y} - E[B\vec{y}])(B\vec{y} - E[B\vec{y}])^T]$$

dove $\vec{\hat{\theta}}$ è il vettore delle stime dei parametri. Da cui:

$$V_{\hat{\theta}} = E[B(\vec{y} - E[\vec{y}])(\vec{y} - E[\vec{y}])^T B^T] = BE[(\vec{y} - E[\vec{y}])(\vec{y} - E[\vec{y}])^T]B^T$$

Per definizione, il valore di aspettazione scritto è la matrice delle covarianze del campione. Quindi:

$$V_{\hat{\theta}} = BVB^T$$

Si può verificare questo risultato anche calcolando esplicitamente le covarianze con la legge di propagazione della varianza. Abbiamo visto che la covarianza tra due generiche funzioni u_i e u_j si ottiene con i loro sviluppi in serie. Fermandosi al primo ordine:

$$cov(u_i, u_j) \approx \sum_{k,l} \frac{\partial u_i}{\partial x_k}\Big|_{\vec{x}=\vec{\mu_x}} \frac{\partial u_j}{\partial x_l}\Big|_{\vec{x}=\vec{\mu_x}} cov(x_k, x_l)$$

Se supponiamo che le u_i siano funzioni lineari delle variabili casuali del campione x_i:

$$\vec{u} = \vec{c} + S\vec{x}, \qquad u_i = c_i + \sum_k S_{ik} x_k$$

dove c_i sono delle costanti e S è una matrice che viene applicata alle x_i. Allora lo sviluppo in serie diventa:

$$cov(u_i, u_j) \approx \sum_{k,l} S_{ik}\, S_{jl}\, cov(x_k, x_l) = \sum_{k,l} S_{ik}\, S_{jl}\, (V_x)_{kl} = (S \cdot V_x \cdot S^T)_{i,j}$$

Ora, le u sono le nostre stime con termine costante nullo. Quindi effettivamente i due risultati coincidono. □

Per quanto riguarda le varianze delle stime, possiamo calcolare:

$$\begin{aligned} BVB^T &= (A^TV^{-1}A)^{-1}A^TV^{-1}\cancel{V}\cancel{V^{-1}}A(A^TV^{-1}A)^{-1} = (A^TV^{-1}A)^{-1} \cdot A^TV^{-1}A \cdot (A^TV^{-1}A)^{-1} = \\ &= (A^TV^{-1}A)^{-1} \cdot I = (A^TV^{-1}A)^{-1} \end{aligned}$$

dove il risultato trovato è molto più semplice da calcolare rispetto all'espressione dalla quale siamo partiti. Inoltre, si è riusciti a trovare un'espressione esatta.

Un'altra cosa interessante di questo metodo e di cui abbiamo già accennato è il fatto che, siccome le stime sono combinazioni lineari del campione, esse ereditano la funzione di distribuzione del campione. In particolare, se questa è gaussiana, anche le stime avranno distribuzione normale. La funzione di distribuzione congiunta è:

$$f(x_1, \dots, x_n) = \frac{1}{(2\pi)^{\lambda/2} |det\ V_\theta|^{1/2}}\, e^{-Q^2/2} \qquad \text{dove} \qquad Q^2 = (\vec{\hat{\theta}} - \vec{\theta}) V_\theta^{-1} (\vec{\hat{\theta}} - \vec{\theta})$$

cioè Q^2 è la forma quadratica che compare nella multinormale. Per il MMQ lineare si trova che vale:

$$X^2 = X_{min} + Q^2 \tag{5}$$

Quindi, se voglio trovare i punti corrispondenti a una deviazione standard, la condizione da porre è:

$$X^2 = X^2_{min} + 1$$

Dimostrazione. Ora vogliamo dimostrare la (5). Si fa la seguente sostituzione:

$$\vec{y} - A\vec{\theta} = \vec{y} - A\vec{\theta} + A\vec{\hat{\theta}} - A\vec{\hat{\theta}} = \vec{y} - A\vec{\hat{\theta}} + A(\vec{\hat{\theta}} - \vec{\theta})$$

e sostituire nella forma quadratica X^2:

$$\begin{aligned} X^2 &= (\vec{y} - \vec{\mu})^T\, V^{-1}\, (\vec{y} - \vec{\mu}) = \\ &= (\vec{y} - A\vec{\hat{\theta}})^T V^{-1} (\vec{y} - A\vec{\hat{\theta}} + A(\vec{\hat{\theta}} - \vec{\theta})) + (\vec{\hat{\theta}} - \vec{\theta})^T V^{-1} A^T (\vec{y} - A\vec{\hat{\theta}} + A(\vec{\hat{\theta}} - \vec{\theta})) = \\ &= (\vec{y} - A\vec{\hat{\theta}}) V^{-1} (\vec{y} - A\vec{\hat{\theta}}) + (\vec{y} - A\vec{\hat{\theta}})^T V^{-1} A(\vec{\hat{\theta}} - \vec{\theta}) + (\vec{\hat{\theta}} - \vec{\theta})^T A^T V^{-1} (\vec{y} - A\vec{\hat{\theta}}) + \\ &+ (\vec{\hat{\theta}} - \vec{\theta})^T A^T V^{-1} A(\vec{\hat{\theta}} - \vec{\theta}) \end{aligned}$$

in cui riconosciamo un po' di termini:

$$(\vec{\hat{\theta}} - \vec{\theta})^T A^T V^{-1} A(\vec{\hat{\theta}} - \vec{\theta}) = (\vec{\hat{\theta}} - \vec{\theta})^T V_\theta (\vec{\hat{\theta}} - \vec{\theta}) = Q^2$$

mentre il secondo e il terzo termine sono nulli (cfr. la derivata che abbiamo fatto all'inizio). Da cui:

$$X^2 = X^2_{min} + Q^2$$

□

La (5) è il risultato a cui si accennava nell'introduzione a questo metodo di stima dei parametri. È esatto in questo caso ma ha comunque validità generale sotto forma di approssimazione.

Ora vediamo alcune applicazioni del MMQ lineare.

Esempio 1: Abbiamo un campione di n variabili indipendenti $y_1, \dots, y_n$, dove $y = mx + q$. Supponiamo che in corrispondenza di ciascuna y_i possiamo associare un'incertezza statistica σ_i^2 e abbiamo calcolato la x_i. Vogliamo utilizzare il campione per stimare i parametri m e q. Potremmo scrivere la forma quadratica X^2 e minimizzarla, ma in alternativa possiamo usare il MMQ lineare.

$$\vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \vec{\theta} = \begin{pmatrix} m \\ q \end{pmatrix}, \qquad A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}, \qquad V = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

Quindi:

$$B = (A^T V^{-1} A)^{-1} A^T V^{-1}$$

Ci sono un po' di calcoli da fare ma si ottiene:

$$\begin{pmatrix} \hat{m} \\ \hat{q} \end{pmatrix} = B \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

che sono effettivamente, scrivendo esplicitamente le matrici, le funzioni delle sommatorie S_{00}, S_{01}, S_{10} e S_{11} che abbiamo visto all'inizio del corso. Inoltre:

$$(A^T V A)^{-1} = V_{\hat{\theta}}$$

Esempio 2: Supponiamo di avere due esperimenti che hanno misurato la grandezza fisica x. Nel primo abbiamo trovato il risultato $x_1 \pm \sigma_1$, nel secondo $x_2 \pm \sigma_2$, indipendenti l'una dall'altra perché gli esperimenti sono totalmente diversi. Poiché la grandezza è la stessa, il valore di aspettazione è lo stesso μ_x. Nel secondo esperimento in più si è misurata un'altra grandezza $z \pm \sigma_z$ con valore di aspettazione μ_z (z è una variabile correlata con x, quindi il coefficiente di correlazione con esso è diverso da zero). Vogliamo combinare gli esperimenti per trovare la migliore stima per μ_x. Supponendo che non ci interessi per il momento l'informazione su z:

$$\vec{y} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad E[\vec{y}] = A\vec{\theta} = \begin{pmatrix} \mu_x \\ \mu_x \end{pmatrix}, \qquad A = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad V = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

Vediamo che la matrice A assume una forma speciale. Applicando il MMQL, troviamo che la migliore stima del valore di aspettazione è, in questo caso, la media pesata di x_1 e x_2:

Se, invece, considerassimo l'informazione su z:

$$\vec{y} = \begin{pmatrix} x_1 \\ x_2 \\ z \end{pmatrix}, \quad E[\vec{y}] = \begin{pmatrix} \mu_x \\ \mu_x \\ \mu_z \end{pmatrix} = A\vec{\theta}, \quad \vec{\theta} = \begin{pmatrix} \mu_x \\ \mu_z \end{pmatrix} \quad A = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad V = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & \rho\sigma_z\sigma_2 \\ 0 & \rho\sigma_z\sigma_2 & \sigma_z^2 \end{pmatrix}$$

Anche in questo caso, si trova una media pesata come migliore stima di μ_x, quindi l'informazione su z non aggiunge nulla. Per μ_z, invece, si trova:

$$\mu_z = z + \rho \frac{\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2}(x_1 - x_2)$$

Il che ha senso: se trovo lo stesso valore di x_1 e x_2, μ_z dipende unicamente da z. Inoltre, la stima della varianza di z è:

$$\sigma^2_{\hat{\mu}_z} = \sigma_z^2 \left(1 - \rho^2 \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)$$

quindi, se i valori di x_1 e di x_2 sono correlati, la varianza stimata risulta essere minore rispetto a quella associata a z.

6 Il test di ipotesi

Abbiamo evidenziato un'importante criticità all'interno della stima dei parametri. Infatti, quest'ultima fornisce, come metodo, sempre la loro stima migliore anche se le ipotesi e le assunzioni di partenza sono manifestamente false. Abbiamo fatto l'esempio di un fit lineare effettuato su dei punti sperimentali che palesemente non seguivano tale andamento. Qua entra in gioco il test di ipotesi. Esso, infatti, serve quando si vuole verificare e quantificare la bontà delle assunzioni fatte. Il problema del test di ipotesi è: come utilizzare al meglio il campione a disposizione per verificare la correttezza o meno dell'ipotesi? Si associa una certa probabilità di errore nell'assumere che un'ipotesi sia sbagliata oppure corretta, sulla base del campione finito a disposizione.

Il test opera su un'ipotesi statistica, che può essere una funzione di distribuzione, una relazione tra grandezze fisiche, ecc...Di seguito sono riportati i passaggi principali di questo metodo di verifica delle ipotesi:

1. Innanzitutto, deve essere ben definita in modo molto chiaro l'ipotesi statistica (ipotesi nulla H_0) che si vuole verificare.

2. Si introduce la statistica di test che è una statistica, ovvero una funzione del solo campione $t(x_1, \ldots, x_n)$ casuale, che è a sua volta una variabile casuale (perché cambia cambiando il campione) con una sua funzione di distribuzione $\phi_0(t)$. Bisogna conoscere quest'ultima distribuzione supponendo che H_0 sia corretta.

3. Si fissa (cioè si definisce a priori) il livello di significatività del test α.

4. Si introduce la regione critica costituita dai valori $t > t_\alpha$ dove

$$t_\alpha : \int_{t_\alpha}^{\infty} \phi_0(t)dt = \alpha.$$

   ovvero è il valore della statistica di test tale per cui l'area sottesa alla curva, da esso alla fine del dominio della statistica, è proprio il livello di significatività fissato.

5. Se il risultato della statistica di test cade nella regione critica, si rigetta H_0. In altre parole, tale regione è "critica" perché se la statistica di test assume un certo valore t^* che cade in essa ($t^* > t_\alpha$), allora si rifiuta H_0.

6. Il test di ipotesi non è significativo $t < t_\alpha$ perché non è possibile stabilire se l'ipotesi sia vera oppure falsa (non si hanno sufficienti elementi per giudicarla). In altre parole, se si sceglie $\alpha = 5\%$, allora non si può dire che l'ipotesi sia corretta con una probabilità del 95%.

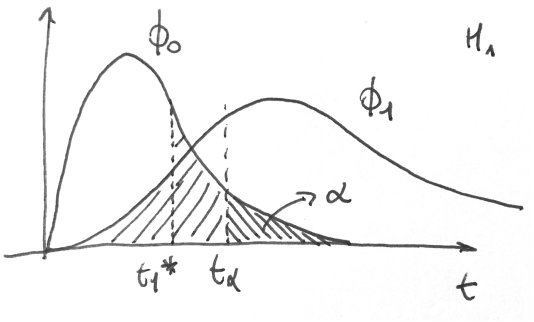
Nota: nulla impedisce che si rigetti un'ipotesi corretta (errore del primo tipo). Per questo motivo, quando si effettua un test di ipotesi, è importante indicare il livello di significatività con il quale si è lavorato, perché è proprio la probabilità di errore nel valutare la correttezza o meno dell'ipotesi assunta.

Naturalmente, si cerca di minimizzare questo tipo di errori. Tuttavia, spesso se ne presenta un secondo tipo, più sottile, che consiste nell'accettare l'ipotesi nulla H_0 anche se è vera l'ipotesi alternativa H_1. Se le ipotesi sono, per esempio, delle funzioni di distribuzione, questo vorrebbe dire che la stessa variabile ne seguirebbe due diverse, il che è assurdo. Questo tipo di errore è legato a valori scelti di α troppo bassi perché, in tal caso, sarebbe bassa la probabilità di rigettare H_0 anche se si trova che l'ipotesi alternativa H_1 è valida.

Nota: Nel test di ipotesi non si sceglie α troppo piccolo perché altrimenti la probabilità di valutare come corretta un'ipotesi di test sbagliata diventa molto alta.

— **Fine nota**

Quando si aggiunge un'ipotesi alternativa, bisogna stare molto attenti a come si sceglie il livello di significatività perché se aumenta la probabilità che la prima ipotesi sia vera, diminuisce quella relativa alla seconda (il test di ipotesi, in questo modo, diventerebbe inutile). Allora, bisogna trovare un buon compromesso. La scelta della statistica di test è estremamente importante



perché deve essere in grado di minimizzare gli errori del primo e del secondo tipo.

La generalità del test di ipotesi permette di applicarlo in moltissime contesti diversi.

In alternativa a usare il livello di significatività α, si introduce il "p-valore":

$$p = \int_{t^*}^{\infty} \phi_0(t)dt$$

dove $t*$ è il valore ottenuto per la statistica di test su uno specifico campione. In questo caso, questo parametro fornisce la probabilità che la statistica assuma un valore maggiore di quello che ha assunto sul particolare campione. Conseguentemente, minore è il p-valore, minore è l'intervallo di integrazione considerato e minore è la probabilità di rigettare l'ipotesi assunta.

Esistono diversi tipi di test:

1. Parametrici: test specifici sui valori dei parametri delle funzione di distribuzione (solitamente distribuzione di Gauss), cioè valore di aspettazione e varianza principalmente, molto diffuso.
2. Non parametrici
3. A una coda
4. A due code: la regione critica viene divisa in due regioni (intervalli di valori), una contenente valori troppo piccoli e l'altra contenente valori troppo alti:

   $$t < t_{1,\alpha/2}, \qquad t > t_{2,\alpha/2}$$

   Si scelgono in modo tale che:

   $$\int_{t_{min}}^{t_{1,\alpha/2}} \phi_0(t)dt = \frac{\alpha}{2} = \int_{t_{2,\alpha/2}}^{t_{max}} \phi_0(t)dt.$$

6.0.1 Il test di ipotesi di χ^2

Sia H_0 l'ipotesi nulla che afferma che delle variabili casuali $Y_1, \ldots, Y_n$ seguono una distribuzione normale $N(\mu_i, \sigma_i^2)$ e con valori di aspettazione $\mu_1, \ldots, \mu_n$. Siano $y_1, \ldots, y_n$ un campione a disposizione delle variabili, per esempio i risultati delle misure, con distribuzione normale. Come

statistica di test si può introdurre la seguente sommatoria:

$$t = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma_i^2}$$

che, se H_0 è corretta e, quindi, le y_i seguono effettivamente una distribuzione normale e i valori di aspettazione indicati sono giusti, allora è una variabile di χ^2_n. In merito, i valori di t_α una volta fissato α si trovano tabulati.

Considerazione #1: Se per caso si trova un basso valore per la statistica di test t, vuol dire che $y_i - \mu_i$ è piccolo, ovvero la differenza tra i valori misurati e i corrispondenti valori di aspettazione è piccola in termini di deviazione standard. Di conseguenza, proprio perché y_i e μ_i si avvicinano come valori, anche in tal caso non c'è motivo di dubitare della veridicità di un'ipotesi. Dal momento che valori bassi della statistica di test non creano problemi, i test di χ^2 sono tipicamente test a una coda.

Ci sono due possibilità in merito all'ipotesi nulla H_0:

1. Ho un campione di dati a partire dal quale si formula l'ipotesi.

   Esempio: scarica condensatore, abbiamo i nostri dati che possiamo mettere in grafico tensione-tempo, possiamo ipotizzare che abbiano un andamento esponenziale.

2. Faccio l'ipotesi a priori e costruisco un campione che permetta di verificarla.

Considerazione #2: Se le variabili del campione sono correlate (covarianze non nulle), allora la statistica di test che si usa è una generalizzazione di quella usata per variabili indipendenti:

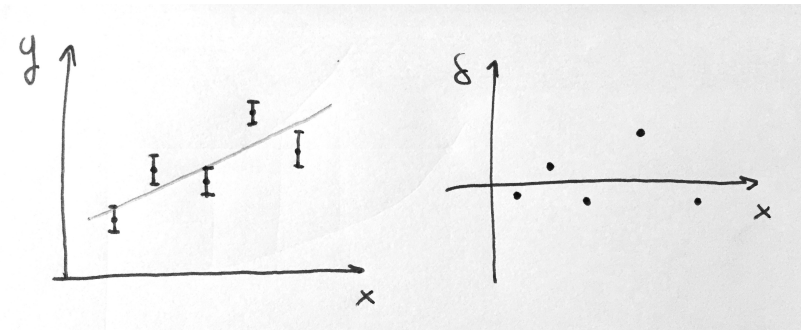
$$t = (\vec{y} - \vec{\mu})^T V^{-1} (\vec{y} - \vec{\mu})$$

La variabile t segue ancora una distribuzione di χ^2_n.

Considerazione #3: La distribuzione di χ^2_n ha valore di aspettazione pari a n. Di conseguenza, ciascun termine contribuisce in media con 1:

$$\delta = \frac{y_i - \mu_i}{\sigma_i} \approx 1$$

In media, tutti i punti devono trovarsi a una deviazione standard. Ma non deve essere uno necessariamente per tutti i punti. Guardando i residui:



In generale, questa è una cosa che è sempre bene controllare perché il valore assunto dalla statistica di test è molto significativo e contiene informazioni importanti. Per esempio, se per caso l'ipotesi fatta non è corretta, si vede che nel grafico dei residui i punti seguono un andamento lineare fino a un certo punto e dopo cominciano a discostarsi. Questo discostamento, contenuto nei singoli contributi alla statistica di test, è un campanello d'allarme e consente di andare a rivedere la correttezza delle ipotesi fatte.

Considerazione #4: La distribuzione di χ^2_n ha varianza pari a $2n$. Se $n \gg 1$, ovvero il campione a disposizione è molto grande, le tabelle chi si possono trovare forniscono il valore degli integrali fino a $n \approx 100$. Infatti, per $n \to \infty$, la funzione di distribuzione di χ^2 tende a una distribuzione normale $N(n, 2n)$. Dunque, per quei valori di n non servono le tabelle di χ^2 ma si può fare affidamento a quelle sulla distribuzione normale. Questo fatto è molto utile perché si può prevedere come vanno le cose senza dover necessariamente consultare le tabelle relative alla distribuzione di χ^2_n.

Esempio: Supponiamo di avere $n = 200$ misure. Quindi, $\sigma_t = \sqrt{2n} = 20$ e la statistica di test attesa è $t^* = 200$. Se invece abbiamo $t^* = 220$, ci troviamo a una deviazione standard e quindi l'integrale "fuori" (cioè il $p - valore$) è del 16%. Se abbiamo $t^* = 240$ ci troviamo a due deviazioni standard e l'integrale è dell'ordine dello 0.025. Se avessi fissato un livello di significatività del 5%, avrei dovuto scartare questo valore per esempio. Se infine $t^* = 160$, il p valore sarebbe dell'ordine di 0.975. Il test non sarebbe significativo, ma un simile risultato dovrebbe insospettirci: forse abbiamo sovrastimato gli errori da associare alle misure.

Considerazione #5: Se c'è una relazione tra le variabili (non sono indipendenti), la distribuzione è sempre di χ^2 ma cambia il numero di gradi di libertà: non è più n (cioè il numero totale di misure) ma $n - r$. In particolare, se avessi usato lo stesso campione per stimare parametri legati a un'ipotesi nulla H_0, equivale ad avere r relazioni tra le misure.

Considerazione #6: Quando si usa il MMQ, si ha già la statistica di test, che è la forma quadratica.

Considerazione #7: Questo tipo di test è poco sensibile al segno (contributi positivi e negativi spesso vanno confusi perché compaiono dei quadrati) e non è uno dei test migliori, anche se è semplice e molto efficiente. Vediamo degli esempi specifici di applicazione del test di χ^2.

Esempio 1: testare la relazione tra due grandezze fisiche.

Abbiamo due grandezze fisiche X e Y legate tra loro:

$$Y = Y(X)$$

Abbiamo un campione di queste due grandezze di dimensione n costituito da coppie x_i, y_i dove le x_i sono note senza errori, mentre le y_i hanno distribuzione gaussiana (perché, per esempio, l'errore statistico è dovuto a errori accidentali) con varianza σ_i^2 nota. Sono misure indipendenti

Possiamo supporre (ipotesi nulla) che $Y = mX + q$, cioè che la relazione sia lineare attraverso due parametri che possono essere noti (numero di gradi di libertà pari a n) oppure sono da determinare dagli stessi dati (numero di gradi di libertà pari a $n - 2$).

Supponiamo di essere nel primo caso. Allora il valore di aspettazione delle y_i è:

$$\mu_i = E[y_i] = mx_i + q$$

La statistica di test è:

$$t = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma_i^2}$$

e quindi fissare il livello di significatività α e stabilire la regione critica:

$$\alpha = 0.05 \to t_\alpha$$

e vedere se rigettare o meno l'ipotesi nulla iniziale. Avevamo fatto qualcosa del genere nell'esercitazione 3, in cui avevamo simulato molte volte un esperimento e usato su ciascuno il MMQ per stimare i parametri:

$$n = 7, \qquad \nu = 5, \qquad t_{\alpha=0.05} = 11$$

dove n è il numero totale di misure e ν il numero di gradi di libertà. Si ha la regione critica per $t^* > 11$. In questo caso i valori dei parametri erano stimati. Se per caso fosse stato noto, si avrebbe avuto:

$$n = 7, \qquad \nu = 7, \qquad t_{\alpha=0.05} = 14$$

prestando particolare attenzione al numero di gradi di libertà e la regione critica, questa volta, si ha per $t^* > 14$.

Infine, si possono mettere in istogramma i valori della statistica di test, verificare che segua un andamento di χ^2_ν con un certo numero consistente di gradi di libertà e che proprio il 5% viene escluso dal grafico.

Esempio 2: verificare l'ipotesi su una sola variabile casuale (test di χ^2 di Pearson)

Sia $x_1, \ldots, x_n$ un campione di dimensione n di variabili casuali che seguono funzione di distribuzione $f(x)$, secondo l'ipotesi nulla H_0. Conosciamo il valore di aspettazione e la varianza di x. Come statistica di test, non è ottimale usare:

$$t = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma_i^2}$$

Infatti, se abbiamo abbastanza dati, possiamo istogrammare i valori e, dunque, passare a un campione $n_1, \ldots, n_m$ di dimensione minore ($m < n$) a quello di partenza e corrispondente agli eventi in ciascuna colonna dell'istogramma. Questo secondo campione ha distribuzione multinomiale però, se il numero di intervalli è sufficientemente grande (probabilità che evento cada in un certo intervallo è molto minore di 1, massimo 20%), allora si possono trascurare le covarianze e considerare gli eventi in ciascun intervallino come indipendenti. Inoltre, la distribuzione degli eventi può essere considerata come una di Poisson invece che binomiale. Per la distribuzione di Poisson, la varianza è uguale al valore di aspettazione.

Conoscendo la funzione di distribuzione dei conteggi, la probabilità che il singolo evento cada in un determinato intervallo di punto centrale x^* e ampiezza Δ:

$$p_i \approx f(x_i^*)\Delta$$

Il valore di aspettazione riferito al singolo intervallo è:

$$\mu_i = E[n_i] = np_i$$

E quindi posso considerare come statistica di test:

$$t = \sum_{i=1}^{m} \frac{(n_i - \mu_i)^2}{\mu_i}$$

Assumendo come vera l'ipotesi nulla, questa è una quantità che si può calcolare e che assume un valore specifico su un campione specifico.

Per eseguire il test di ipotesi devo conoscere la funzione di distribuzione di t. Se gli n_i hanno funzione di distribuzione normale $N(\mu_i, \mu_i)$ (il che è possibile perché Poisson diventa normale quando n diventa grande, cioè $n \geq 10$), allora t ha una distribuzione di χ^2.

Esempio 3: compatibilità dei risultati

Supponiamo che x sia una grandezza fisica di cui si conosca il valore, per esempio μ. Abbiamo fatto un esperimento e abbiamo trovato un valore $x_1 \pm \sigma_1$, ovvero con un'incertezza statistica associata. Essendo gli errori di natura accidentale, è ragionevole supporre che x_1 segua una distribuzione normale $N(\mu, \sigma_1^2)$. Vogliamo vedere se il valore trovato x_1 è in accordo o meno con il valore previsto μ. Dunque, l'ipotesi nulla da fissare è:

$$H_0 : E[x_1] = \mu$$

Si introduce la statistica di test:

$$t = \frac{(x_1 - \mu)^2}{\sigma_1^2}$$

che ha distribuzione di χ_1^2, se l'ipotesi nulla è corretta, ovvero se x_1 segue una distribuzione normale. Successivamente, si introduce il livello di significatività. Ne consegue:

$$t_\alpha : \int_{t_\alpha}^{\infty} f(t)dt = \alpha.$$

Sulle tavole si trova che $t_\alpha = 3.84$ quando $\alpha = 0.05$. Se il valore di t calcolato nell'esperimento è maggiore di t_α, rigetto l'ipotesi. In tal caso, il risultato x_1 non risulta essere compatibile con il valore previsto μ.

Questo test è del tutto equivalente all'introduzione di una statistica di test diversa:

$$t' = \frac{x - \mu}{\sigma_1}$$

cioè la radice di quella di prima. Questa statistica non ha più distribuzione di χ_1^2 ma una normale standard $N(0, 1)$, che conosco e si trova tabulata. Anche in questo caso si può fare il test di ipotesi, scegliendo lo stesso livello di significatività, ma in questo caso non andrà bene lavorare con numeri troppo piccoli o troppo grandi rispetto a μ. Di conseguenza, con questa statistica conviene lavorare con un test a due code, definendo come regione critica:

$$t' < -t'_{\alpha/2}, \qquad t' > t'_{\alpha/2}$$

cioè è costituita da due intervalli. Si trova tabulato $t'_{\alpha/2} = 1.96$ quando si pone $\alpha = 0.05$. Riassumendo, il test di ipotesi così svolto rappresenta la procedura da adottare se si vogliono analizzare in modo del tutto quantitativo i risultati di un esperimento e non, invece, in modo qualitativo usando le regioni definita da una, due e tre deviazioni standard.

Se invece di avere una sola misura dall'esperimento ne abbiamo due, $x_1 \pm \sigma_1$ e $x_2 \pm \sigma_2$, ed è noto il valore di aspettazione di entrambe (è lo stesso), allora è la cosa più semplice è usare un test di ipotesi di χ^2. La statistica di test sarà:

$$t = \frac{(x_1 - \mu)^2}{\sigma_1^2} + \frac{(x_2 - \mu)^2}{\sigma_2^2}$$

che ha una distribuzione di χ_2^2 se si suppone che x_1 e x_2 seguano una distribuzione normale. Se si considera $\alpha = 0.05$, si trova tabulato che $t_\alpha = 5.99$. I valori per la statistica di test superiori a quest'ultimo valore vanno scartati. Se, invece, si considera $\alpha = 0.10$, $t_\alpha = 4.61$.

Nel caso di due misure, potrei considerare anche la media pesata come loro valore stimato:

$$\hat{\mu} = \frac{\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$

e usare come statistica di test:

$$t = \frac{\hat{\mu} - \mu}{\sigma_{\hat{\mu}}}$$

che ha distribuzione normale standard se x_1 e x_2 hanno distribuzione normale. Il test di ipotesi diventa a due code, si determina la regione critica ed eventualmente si scarta l'ipotesi nulla. Questo approccio è equivalente al precedente.

Nel caso generale di n misure $x_1, \ldots, x_n$ con distribuzione gaussiana, se conosco il valore di aspettazione posso scrivere la statistica di test come:

$$t = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma_i^2}$$

che segue una distribuzione di χ^2_n. Se invece non lo conosco, cioè devo stimarlo dai dati stessi, posso usare come statistica di test:

$$t = \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{\sigma_i^2}, \qquad \hat{\mu} = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}$$

che segue una funzione di distribuzione di χ^2_{n-1}.

Se le misure non sono indipendenti, ovvero hanno covarianza non nulla, si prende come statistica di test:

$$t = (\vec{x} - \vec{\mu})^T V^{-1} (\vec{x} - \vec{\mu}) \qquad \text{dove} \qquad \vec{\mu} = (\mu, \ldots, \mu)$$

che ha una funzione di distribuzione di χ^2_n nel caso in cui si conosca il valore di aspettazione e di χ^2_{n-1} altrimenti.

Ora supponiamo di avere due variabili casuali x_1 e x_2. Invece di costruire la statistica di test:

$$t = \frac{(x_1 - \hat{\mu})^2}{\sigma_1^2} + \frac{(x_2 - \hat{\mu})^2}{\sigma_2^2}$$

che ha distribuzione χ^2_1, posso usare la differenza tra i due valori e fare il test di ipotesi con zero, ovvero vedere quanto la differenza è compatibile con zero:

$$\delta = x_1 - x_2, \qquad t = \frac{(\delta - \mu_\delta)^2}{\sigma_\delta^2} = \frac{\delta^2}{\sigma_\delta^2} = \frac{(x_1 - x_2)^2}{\sigma_1^2 + \sigma_2^2}$$

che ha funzione di distribuzione di χ^2_1. Oppure posso usare la statistica di test:

$$t' = \frac{x_1 - x_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

che ha funzione di distribuzione normale standard.

Se le due variabili non sono indipendenti, la varianza sulla differenza non è più la semplice somma delle varianze sulle singole variabili ma ha un termine aggiuntivo (secondo le legge di propagazione della varianza nel caso generale):

$$\sigma_\delta^2 = \sigma_1^2 + \sigma_2^2 + 2 \frac{\partial \delta}{\partial x_1} \frac{\partial \delta}{\partial x_2} \operatorname{cov}(x_1, x_2) = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$$

dove ρ è il coefficiente di correlazione. Quindi, per esempio, si può stimare in due modi diversi lo stesso parametro e vedere quanto la sua differenza sia compatibile con zero. La differenza

non dovrebbe essere grande perché altrimenti vorrebbe dire che uno dei due metodi non sarebbe valido. Inoltre, mi aspetto che, se i due metodi sono corretti, le due stime siano correlate positivamente (se il campione suggerisce un valore del parametro maggiore del valore di aspettazione, allora ciò si manifesta nella stima indipendentemente dal metodo adottato per calcolarla), ovvero $\rho = 1$ e di conseguenza $\sigma_\delta^2 \approx \sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2 = (\sigma_1 - \sigma_2)^2$ con $\delta = \theta_1 - \theta_2$ La statistica di test è:

$$t = \frac{(\theta_1 - \theta_2)^2}{(\sigma_1 - \sigma_2)^2}$$

che segue una distribuzione di χ_1^2. La cosa importante è che al denominatore ho la differenza tra le deviazioni standard, che generalmente è molto minore della deviazione standard della singola stima. Infatti, per verificare la compatibilità tra due misure è necessario non solo confrontare con le singole deviazioni standard, ma anche con la loro differenza.

Esempio 4: test di indipendenza tra grandezze fisiche caratteristiche

Questa particolare applicazione del test di χ^2 serve per testare la dipendenza o indipendenza tra caratteristiche diverse di una popolazione. Supponiamo di avere un campione di n persone e di voler testare che altezza X e il peso Y siano due caratteristiche indipendenti, ovvero se i valori che assumono siano correlati o meno. La funzione di distribuzione congiunta di X e Y, se le due sono indipendenti, è semplicemente il prodotto delle singole fz. di distribuzione:

$$f(X, Y) = f_x(X) \cdot f_y(Y)$$

Si dividono i valori di X e Y in classi:

$$X_1 : h < 1.65, \qquad X_2 : 1.65 < h < 1.75, \qquad X_3 : h > 1.75, \qquad Y_1 : p < 65, \qquad Y_2 : p > 65$$

A questo punto si costruisce una tabella in cui si riporta il numero di elementi in ciascuna classe definita:

	X_1	X_2	X_3	$tot\ riga:$
Y_1	n_{11}	n_{12}	n_{13}	n_{1Y}
Y_2	n_{21}	n_{22}	n_{23}	n_{2Y}
$tot\ colonna:$	n_{X1}	n_{X2}	n_{X3}	n_{XY}

Se X e Y sono indipendenti (questa è la nostra ipotesi nulla):

$$\mu_{ij} = E[n_{ij}] = np_{ij} = n(p_{iY} \cdot p_{jX}) = n\frac{n_{iY}}{n}\frac{n_{jX}}{n} = n_{iY} \cdot n_{jX}\frac{1}{n}$$

dove p_{ij} è la probabilità che l'elemento appartenga alla i-esima classe di Y e alla j-esima classe di X. Si è utilizzata la definizione frequenzista di probabilità per procedere con i calcoli. Se si assume che la probabilità associata al singolo evento sia piccola, posso usare Poisson e affermare che:

$$\sigma_{ij}^2 = \mu_{ij}$$

Inoltre, se ho valori di aspettazione alti (maggiori di 5), allora Poisson→Gauss e posso introdurre la statistica di test:

$$t = \sum_{ij} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$$

che ha una distribuzione di χ_ν^2 con ν il numero di gradi di libertà, legato al numero di caselle (più precisamente al numero di righe e a quello di colonne). Il numero di gradi di libertà non è pari

al numero di caselle. Ci sono infatti una "relazione" nelle righe e una "relazione" nelle colonne: in entrambi i casi, l'ultima probabilità si trova sapendo che la somma di tutte le probabilità è 1. Quindi, bisogna togliere un'unità al numero di righe n_r e un'altra unità al numero di colonne n_c:

$$\nu = (n_r - 1)(n_c - 1) \tag{6}$$

Nota: il test di indipendenza poteva essere applicato all'esercizio in cui abbiamo stimato molte volte coefficiente angolare e intercetta relativa a dei dati. Il grafico $\hat{m} - \hat{q}$ poteva essere suddiviso in tante caselle attraverso una griglia, associare a ciascuna casella una classe di $\hat{m}$ e di $\hat{q}$.

— **Fine nota**

Esempio: Il test di indipendenza può essere usato per testare la correlazione tra genere e attività, ovvero tra il sesso di una persona e il lavoro che svolge. Supposto che ci siano un totale di 240 persone, di cui 100 donne e 140 uomini, la tabella che si costruisce è la seguente: La

	Agricoltura	Artigianato	Industria	Servizi
Donne	0	8	12	80
Uomini	10	52	58	20

probabilità relativa a Y_1, ovvero alla classe "donne" è di 0.42. Quella relativa a Y_2, ovvero alla classe "uomini" è di 0.58. Le probabilità relative alle classi X_1, X_2, X_3, X_4, ovvero quelle delle classi lavorative, sono rispettivamente di 0.04, 0.25, 0.29 e di 0.42. Facendo il prodotto delle probabilità, si ricavano quelle relative a ciascuna casella della tabella. I valori attesi si ottengono moltiplicando per il numero totale della popolazione (n=240): A questo punto, si può calcolare

	Agricoltura	Artigianato	Industria	Servizi
Donne	4.2	25.0	29.2	41.7
Uomini	5.8	35.0	40.8	58.3

la statistica di test che risulta essere uguale a 105, con distribuzione di χ^2_ν dove ν è il numero di gradi di libertà ed è pari a 3 (per la (6)). Se si considera un livello di significatività $\alpha = 0.05$, si trova $t_\alpha = 7$. Quindi, la statistica di test assume un valore decisamente maggiore, cade nella regione critica e si rigetta l'ipotesi che genere e attività siano due caratteristiche indipendenti.

6.0.2 Test parametrici

I test parametrici sono un altro insieme di test d'ipotesi con uso diffuso e sono relativi ai parametri di una funzione di distribuzione che, in genere, è quella normale. Un esempio tipico è un test sul valore atteso quando la varianza è nota (o non lo è). L'ipotesi nulla, in generale, è:

$$H_0 : \mu = \mu_0$$

dove μ è il valore del parametro e μ_0 è un certo valore numerico. Si usano come ipotesi alternative:

$$H_1 : \mu \neq \mu_0 \qquad H_2 : \mu < \mu_0, \qquad H_3 : \mu > \mu_0$$

Per H_0 e H_1 si deve usare un test a due code perché in tale condizione non andrebbero bene né valori troppo grandi né troppo piccoli. Al contrario, per le altre due ipotesi H_2 e H_3 il test è a una coda.

Esempio 1:

Supponiamo di avere n misure $x_1, \ldots, x_n$ che seguono tutte una distribuzione normale $N(\mu_0, \sigma^2)$, dove la varianza è nota. Ipotesi nulla H_0: il valore di aspettazione è μ_0. Considero la media delle misure:

$$\bar{x} = \frac{1}{n} \sum_i x_i$$

Se l'ipotesi nulla è corretta, avrà funzione di distribuzione normale $N\left(\mu_0, \frac{\sigma^2}{n}\right)$. La statistica di test sarà:

$$t = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Se l'ipotesi nulla è corretta, t ha una distribuzione normale standard. Si fissa un livello di significatività pari a $\alpha = 0.05$. Considero due code della distribuzione se devo verificare H_0 oppure H_1, segnando quindi il valore $\frac{\alpha}{2}$. Invece, per le ipotesi H_2 e H_3 considero solo una coda e, di conseguenza, uso direttamente il valore di α.

Esempio 2:

Vogliamo fare un test sulla produzione di confezioni di prodotti alimentari dal peso di 1.6 kg circa ciascuna. Ci aspettiamo che la distribuzione del peso delle confezioni sia gaussiana $N(\mu_0, \sigma^2)$ dove $\mu_0 = 1600\, g$ in teoria e $\sigma = 120\, g$, cioè è nota. Prendiamo 100 confezioni, le pesiamo e troviamo che:

$$\bar{x} = 1570\, g$$

che rappresenta il peso medio. Vogliamo capire se questo risultato è in accordo con il valore nominale (non va bene se vendo una confezione che pesa di meno o di più di quello che dico!). Allora, introduco la statistica di test:

$$t = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = -2.5$$

e fisso il livello di significatività $\alpha = 0.05$, ottenendo $t_\alpha = 1.96$. Dunque, la regione critica è costituita da due intervalli:

$$t < -1.96 \qquad t > 1.96$$

Il valore trovato per t rientra nella prima regione critica, dunque rigetto l'ipotesi nulla H_0 e il peso medio delle confezioni non è compatibile con quello nominale.

Se avessi usato la seconda ipotesi alternativa H_2, cioè se mi preoccupassi di produrre confezioni solo più leggere di quella nominale, non avrei avuto problemi.

Esempio 3:

Se in più non conoscessi la varianza, allora, per eseguire il test sul valore medio, dovrei anche stimare la varianza a partire dal campione $x_1, \ldots, x_n$. Uno stimatore della varianza è:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Quindi, potrei usare questa stima della varianza e procedere come prima, ovvero considerare come statistica di test:

$$t = \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}}$$

In realtà, questa statistica non ha più una funzione di distribuzione normale standard perché è il rapporto tra due variabili casuali (perché al denominatore ho la stima della varianza, che non è un numero ma una variabile casuale vera e propria!). Invece, a meno di un fattore di

scala, la distribuzione è di χ^2_{n-1}. Tuttavia, dividendo numeratore e denominatore per $\sigma/\sqrt{n}$ e moltiplicando e dividendo il numeratore per $n-1$, posso riscrivere la statistica come:

$$t = \frac{\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{S^2}{\sigma^2}\frac{n-1}{n-1}}}.$$

e al denominatore ottengo una variabile con funzione di distribuzione di χ^2_{n-1} divisa per il numero di gradi di libertà.

Si può dimostrare (*) che se una variabile z ed è definita come:

$$z = \frac{x}{\sqrt{y/\nu}}$$

dove x ha una distribuzione normale standard e y ha distribuzione di χ^2_ν, allora z segue la funzione di distribuzione di t-Student, che ha la seguente espressione:

$$h(z) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\cdot\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\sqrt{\nu}}\cdot\left(\frac{z^2}{\nu}+1\right)^{-\frac{\nu+1}{2}}$$

La funzione è massima per $z=0$ ed è simmetrica per altri suoi valori. Per $z\to+\infty$, la funzione di distribuzione tende a zero. Invece, per $\nu\to+\infty$ tende ad assomigliare rapidamente a quella gaussiana. In particolare, per $\nu=1$, la distribuzione è del tipo $1/(1+z^2)$, ovvero è la distribuzione di Cauchy (è come una gaussiana con le code più alte).

Dimostrazione. **(il rapporto tra una distribuzione normale e una di χ^2_ν è una distribuzione di t-Student):**

$$z = \frac{x}{\sqrt{y/\nu}}$$

dove x ha distribuzione normale standard e y ha distribuzione χ^2_ν. Poiché vogliamo calcolare la funzione di distribuzione di z, dobbiamo introdurre una seconda funzione di x e y che ne sia indipendente:

$$z_2 = y$$

Scrivo x in funzione del resto:

$$x = z\sqrt{\frac{y}{\nu}} = z\sqrt{\frac{z_2}{\nu}}$$

Calcolo il determinante dello jacobiano:

$$|J| = \det\begin{pmatrix}\sqrt{\frac{z_2}{\nu}} & \frac{z}{2\sqrt{\nu z_2}}\\ 0 & 1\end{pmatrix} = \sqrt{\frac{z_2}{\nu}} = \sqrt{\frac{y}{\nu}}$$

Quindi, la funzione di distribuzione congiunta è il prodotto delle funzioni di distribuzioni delle singoli variabili:

$$f(x,y) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\cdot\frac{1}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)}y^{\frac{\nu}{2}-1}e^{-y/2}$$

Quindi, sostituendo le espressioni per x e y:

$$g(z,z_2) = \frac{1}{\sqrt{2\pi}2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)}e^{-z^2\cdot\frac{z_2}{2\nu}}\cdot z_2^{\frac{\nu}{2}-1}e^{-z_2/2}\sqrt{\frac{z_2}{\nu}}$$

dove l'ultimo fattore è $|J|$. Quindi:

$$g(z, z_2) = \frac{1}{\sqrt{2\pi} 2^{\nu/2} \Gamma\left(\frac{\nu}{2}\right)} z_2^{\frac{\nu-1}{2}} \exp\left(-\frac{z_2}{2}\left(\frac{z^2}{\nu}+1\right)\right)$$

Quindi, integro questa funzione di distribuzione congiunta:

$$h(z) = \int_0^\infty g(z, z_2)\, dz_2$$

Facendo il cambio di variabile:

$$q = \frac{z_2}{2}\left(\frac{z^2}{\nu}+1\right) \quad \rightarrow \quad dz_2 = \frac{2}{\left(\frac{z^2}{\nu}+1\right)} dq$$

si ottiene, ricordando la definizione della funzione Gamma, che:

$$h(z) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right) \cdot \Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu}} \cdot \left(\frac{z^2}{\nu}+1\right)^{-\frac{\nu+1}{2}}$$

che è proprio la distribuzione di t-Student. □

Esempio 4:
Un altro test che si può fare è quello sulla varianza, nel quale il valore di aspettazione potrebbe essere noto oppure no. Chiaramente, questa seconda possibilità è più complicata.

6.0.3 Relazione tra test di ipotesi e MMQ

Abbiamo a disposizione un campione di n misure y_i di cui si conosce la varianza σ_i^2 e, in una certa ipotesi, il valore di aspettazione è funzione dei parametri che si vogliono stimare, ovvero $E[y_i] = \mu_i(\vec{\theta})$. Se le misure sono indipendenti, il MMQ suggerisce di usare come statistica la forma quadratica:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma_i^2}$$

Il suo valore minimo X^2_{min} è quello corrispondente alle stime dei parametri $\vec{\hat{\theta}}$. Se le y_i hanno distribuzione gaussiana, allora X^2_{min} ha una distribuzione di χ^2_n. Quindi, si può usare direttamente X^2_{min} per fare il test di ipotesi.

Esempio: Ipotizzo un andamento lineare, del tipo: $y = bx + c$ e calcolo la forma quadratica. Se rigetto l'ipotesi, posso farne un'altra aggiungendo un parametro: $y = ax^2 + bx + c$. Cosa non fare assolutamente: aumentare il numero dei parametri fino a quando il valore della forma quadratica è molto basso o addirittura nullo, se no il test è inutile (infatti, in tal caso la relazione supposta sarebbe una polinomiale che passerebbe perfettamente per tutti i punti sperimentali, annullando completamente X^2). L'obiettivo è minimizzare χ^2 in funzione dei parametri, non minimizzare χ^2 aumentando il numero dei parametri.

7 Intervalli di confidenza

Abbiamo visto diverse procedure per stimare i parametri (MMQ e MML) e abbiamo visto che le soluzioni (gli stimatori) possono essere ottenute analiticamente o per metodi numerici. Abbiamo

anche studiato funzioni di distribuzione, in particolare le statistiche di test. Con questo metodo si ottengono le cosiddette "stime puntuali", ovvero il valore stimato del parametro. Tuttavia, queste stime sono a loro volta delle variabili casuali e, dunque, hanno delle funzioni di distribuzione e un'incertezza statistica associate. Quest'ultima può essere determinata tramite la radice quadrata della varianza oppure attraverso gli intervalli di confidenza.

Esempio: Abbiamo un campione di n misure $x_1, \ldots, x_n$ che seguono distribuzione $f(x)$ con valore di aspettazione μ, sia $t(x_1, \ldots, x_n)$ lo stimatore. Si vuole stimare il parametro θ. Si deve avere $\theta = E[x] = \mu$. In questo caso, sappiamo che il migliore stimatore è media aritmetica:

$$t = \frac{1}{n} \sum_i x_i$$

Inoltre sappiamo che se il campione è sufficientemente grande, t segue una distribuzione normale $g(t, \theta) = N(\mu, \sigma_t^2)$ che, dunque, è nota. La probabilità che il valore assunto dallo stimatore rientri nella regione di una deviazione standard è:

$$P(|t^* - \mu| \leq \sigma_t) = 0.68.$$

Quindi, si può dire che l'intervallo $(t^* - \sigma_t, t^* + \sigma_t)$ corrisponde al 68% (livello di confidenza) di probabilità di contenere il valore vero del parametro ed è detto "intervallo di confidenza". Analogamente, si introduce come intervallo $(t^* - 2\sigma_t, t^* + 2\sigma_t)$ relativo al livello di confidenza 95%. Normalmente, le percentuali scelte come livello di confidenza sono piuttosto elevate.

68%	$t^* \pm \sigma_t$
90%	$t^* \pm 1.64\sigma_t$
95%	$t^* \pm 1.96\sigma_t$
99%	$t^* \pm 2.57\sigma_t$

Tabella 7: Intervalli di confidenza Gaussiana

Il discorso è diverso se la funzione di distribuzione dello stimatore non è gaussiana. Un esempio è la stima S^2 della varianza:

$$S^2 = \frac{1}{n-1} \sum_i (x_i - po\bar{o}rx)^2$$

che ha distribuzione di χ^2_{n-1}.

In questi casi, si procede nel modo seguente: per un certo valore del parametro che si vuole stimare, si vuole trovare il valore $u_\alpha = u_\alpha(\theta*)$, che dipende dal particolare valore assunto dal parametro, tale che:

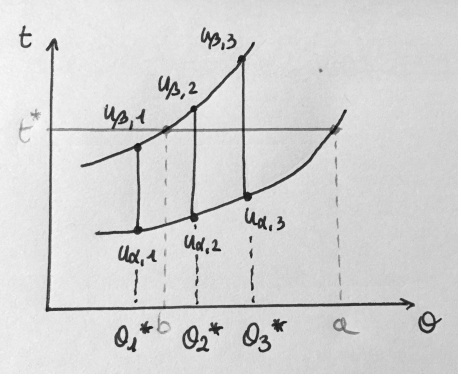
$$P(t \leq u_\alpha) = \int_0^{u_\alpha} g(t, \theta*)dt = \alpha$$

Poi cerco il valore $u_\beta = u_\beta(\theta*)$ tale che:

$$P(t \geq u_\beta) = \int_{u_\beta}^{\infty} g(t, \theta*)dt = \beta$$

Da cui si deduce: $P(t \in [u_\alpha, u_\beta]) = 1 - \alpha - \beta = 1 - \gamma$ ponendo $\gamma = \alpha + \beta$

Se prendo un valore del parametro $\theta*$ diverso, la funzione di distribuzione sarà diversa e quindi cambieranno i valori di u_α e u_β. Questo fatto si può rappresentare graficamente:



La regione compresa tra la curva $u_\alpha(\theta)$ e la curva $u_\beta(\theta)$ individua la "striscia di confidenza" che corrisponde a una probabilità pari a $1-\gamma$ ed è la probabilità che l'intervallo $[b, a]$ in figura, ottenuto fissando un certo valore di t^*, contenga il valore vero del parametro:

$$P(b < \theta < a]) = 1 - \gamma$$

Da notare che quest'ultima affermazione non è equivalente a dire che quella è la probabilità che il valore vero del parametro rientri in quell'intervallo, perché tale valore non è variabile e, dunque, la probabilità associata sarebbe 0 o 1 a seconda dei casi.

Tutto questo è possibile solo se si conosce la funzione di distribuzione delle stime. Inoltre, fissare il valore di $1-\gamma$ non individua in modo univoco l'intervallo $[b, a]$. Per definire gli intervalli di confidenza "centrali" si pone $\alpha = \beta$. Inoltre, alla misura vengono associati degli errori asimmetrici. Questo procedimento si può adattare a più dimensioni.

Ricordiamo che se il parametro ha funzione di distribuzione gaussiana (in generale è vero per campioni di dimensione grande) ci sono i metodi grafici per stabilire l'incertezza: il maximum Likelihood-0.5, i minimi quadrati+1. Se la funzione non è esattamente Gaussiana, ma un po' asimmetrica, possiamo associare errori asimmetrici con questi metodi. Quando la funzione di distribuzione delle stime non è nota, si ricorre, per esempio, a delle tecniche di Monte-Carlo per determinare gli intervalli di confidenza (come abbiamo fatto con il metodo delle repliche nell'esercitazione 5).

7.1 In aggiunta alla lezione sugli intervalli di confidenza

Abbiamo visto diversi metodi per stimare, a partire da un campione di osservazioni $(x_1, ..., x_n)$, il valore di un parametro θ di una funzione $f(x;\theta)$. Il parametro avrà un suo valore vero θ_0, ma ovviamente la stima $\hat{\theta}$ potrebbe risultare diversa da esso avendo a disposizione un campione finito; perciò al valore stimato $\hat{\theta}$ viene attribuita un'incertezza. Il modo più intuitivo di farlo sarebbe quello di voler assegnare una data probabilità al fatto che il valore vero θ_0 sia compreso in un certo intervallo stimato tramite le osservazioni ricordando che $\hat{\theta}$ è a sua volta una variabile casuale con $\sigma^2_{\hat{\theta}}$:

$$P_k = Prob(\hat{\theta} - k\sigma\hat{\theta} \leq \theta_0 \leq \hat{\theta} + k\sigma\hat{\theta})$$

Ma il valore θ_0, sebbene ignoto, è fissato; dunque tale affermazione probabilistica risulta banale: $P_k = 0$ oppure $P_k = 1$! Dunque, dato un particolare campione e la stima $\hat{\theta}$ del parametro θ_0, vorremmo determinare dai dati un intervallo per il quale si possa stimare un probabilità $(1-\gamma)$ che esso contenga il valore vero θ_0:

$$Prob(\hat{\theta}_- \leq \theta_0 \leq \hat{\theta}_+) = (1-\gamma)$$

L'intervallo $[\hat{\theta}_-, \hat{\theta}_+]$ è detto intervallo di confidenza di livello $(1-\gamma)$, dove γ rappresenta la probabilità ("rischio") che l'intervallo non contenga θ_0.

NOTA: La procedura si applica dopo una singola stima θ e dunque deve avere validità indipendente dallo specifico θ_0 che è ignoto.

8 Simulazione di Montecarlo

Vengono usate ad esempio negli esperimenti dell'LHC: confronto tra dati sperimentali raccolti e che cosa prevede la teoria attraverso simulazioni con il metodo di Montecarlo che hanno un'altissima precisione.

Esistono molti metodi per generare numeri casuali oltre, ad esempio, alla subroutine `random_number` di Fortran, usata per avere una distribuzione uniforme tra 0 e 1. Con questi, la distribuzione dei numeri casuali può essere la più varia. I metodi che vediamo noi a lezione sono i seguenti:

- Metodo dell'accept/reject;
- La funzione di ripartizione;
- Il metodo della media;
- La trasformata / il metodo di Box-Mueller.

Studiamoli uno alla volta.

- Metodo dell'accept/reject:

  Supponiamo di voler generare numeri con distribuzione generica $f(x)$ e considero il grafico di quest'ultima in un certo intervallo altrettanto generico $[a, b]$. Successivamente, racchiudiamo la funzione in una scatola di lati gli intervalli $[a, b]$ e $[0, f_{max}]$ rispettivamente situati sull'asse delle ascisse e su quello delle ordinate, dove f_{max} è il massimo valore assunto dalla funzione f, in simboli:

  $$f_{max} = \max_{[a,b]} f$$

  L'algoritmo alla base di questo metodo è il seguente:

  1. Genero un numero casuale r' con distribuzione uniforme tra 0 e 1.
  2. Lo sposto in una distribuzione uniforme tra a e b attraverso la formula:

     $$r_1 = a + (b - a)r'$$

  3. Genero allo stesso modo un secondo numero casuale $r_2 \in [0, \max_{[a,b]} f]$.
  4. Vedo quanto vale la funzione nel primo punto individuato: se $f(r_1) > r_2$, accetto r_1 che diventa il mio x_1.
  5. Ripeto il procedimento ottenendo un insieme di punti $x_1, x_2, x_3, ..., x_n$ che posso inserire in un istogramma. L'andamento sarà proprio quello della funzione di distribuzione.

  Tuttavia, ci sono alcuni problemi con l'efficienza di questo metodo, definita come:[4]

  $$\eta = \frac{N_{acc}}{N_{gen}} \approx \frac{\int_a^b f(x)dx}{(b - a) \cdot \max_{[a,b]} f}$$

  Alcuni esempi di utilizzo di questo metodo:

[4] È il rapporto tra il totale di numeri accettati e il totale dei numeri generati. Dal momento che il numeratore è proporzionale all'area sottesa al grafico, mentre il denominatore è proporzionale all'area del rettangolo in cui è racchiusa la funzione, tale rapporto può essere scritto anche nei termini appena definiti.

1.
$$f(x) = \frac{1 + a\cos x}{2\pi}, \quad x \in [0, 2\pi]$$

L'efficienza è:

$$\eta = \frac{\int_0^{2\pi} \frac{1+a\cos x}{2\pi} dx}{2\pi\left(\frac{1+a}{2\pi}\right)} = \frac{\frac{2\pi}{2\pi}}{1+a} = \frac{1}{1+a} = \begin{cases} 1 & a = 0 \\ \frac{1}{2} & a = 1 \end{cases}$$

f è normalizzata e positiva, da cui:

$$|a| \leq 1$$

2.
$$f(x) = \frac{e^{-x/\tau}}{\tau}$$

è normalizzata: vi applico il metodo dell'accept/reject. Vi sono alcuni problemi: per costruire la "scatola" $[a, b] \times [0, \max_{[a,b]} f]$, devo decidere dove fermarmi sull'asse delle ascisse. Già questo è molto poco efficiente perché man mano che aumenta x, diminuisce l'area sottesa e quindi è poco probabile che un numero venga accettato. Per ovviare a questo problema, si possono costruire tante scatole diverse che diventano sempre più piccole. Tuttavia, persiste la necessità di troncare a un certo punto sull'asse orizzontale. Quindi, questo metodo non può essere applicato a qualsiasi funzione di distribuzione perché non sempre è efficiente.

- Funzione di ripartizione: Data $f : [a, b] \to [0, 1]$ distribuzione di probabilità, la funzione di ripartizione è:
$$F(x) = \int_a^x f(x')dx'$$

Normalmente tale funzione ha un andamento sigmoide e varia tra 0 e 1. Sapendo che vale una sorta di conservazione della probabilità:

$$f(x)dx = g(y)dy$$

Da cui:
$$g(y) = f(x(y)) \left|\frac{dx}{dy}\right|$$

Con $y(x) = F(x)$ otteniamo la funzione di distribuzione di F:

$$g(F) = f(x) \left|\frac{dx}{dF}\right| = f(x) \cdot \frac{1}{f(x)} = 1$$

siccome
$$\frac{dF}{dx} = \frac{d}{dx}\int_a^x dx' f(x') = f(x)$$

Vuol dire che la funzione è uniformemente distribuita nell'intervallo $[0, 1]$ sull'asse delle ordinate. Per generare numeri con distribuzione f si può generare la variabile con distribuzione uniforme $r =: F(x)$. Se F è invertibile in senso analitico (e non numerico),

$$x = F^{-1}(r)$$

Seguirà la distribuzione $f(x)$. Facciamo un esempio pratico: prediamo

$$f(x) = \frac{e^{-x/\tau}}{\tau}.$$

Calcolo la funzione di ripartizione:

$$F(x) = \int_0^x f(x')dx' = \int_0^{x/\tau} e^{-\nu} = -e^{-\nu}\Big|_0^{x/\tau} = 1 - e^{-x/\tau} = r$$

Da cui

$$x = -\tau \ln(1 - r).$$

dove x è distribuito con una funzione di distribuzione f.

L'efficienza è 1 perché per ogni numero casuale generato con distribuzione uniforme, si trova un numero x che segue una distribuzione f: accetto tutti i numeri che genero.

- Metodo della media (per distribuzione normale standard): Per generare dei numeri con distribuzione normale standard, si usa la media di valori generati con distribuzione uniforme:

$$\bar{x} = \sum_{i=1}^{N} \frac{r_i}{N}.$$

dove r_i ha distribuzione uniforme in $[0, 1]\,\forall i$. $\bar{x}$ risulta seguire una distribuzione gaussiana, ma non standard. Infatti, valore di aspettazione e varianza sono rispettivamente diversi da 0 e 1:

$$E[\bar{x}] = \mu = E[\sum_{i=1}^{N} \frac{r_i}{N}] = \frac{1}{N} \sum E(r_i) = \frac{N}{N} 1/2 = 1/2.$$

Mentre la sua varianza è:

$$\sigma_{\bar{x}}^2 = \frac{1}{N^2} \sum \sigma_{r_i}^2 = \frac{1}{N^2} N \frac{1}{12} \implies \sigma_{\bar{x}} = \frac{1}{\sqrt{12N}}$$

Per ottenere la distribuzione normale standard, per il teorema del limite centrale, bisogna considerare:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - 1/2}{\frac{1}{\sqrt{12N}}}$$

che segue la distribuzione $N(0, 1)$. Se, invece, volessimo una variabile z' che segua una distribuzione $N(\mu', \sigma')$, potremmo definire z' nel modo seguente:

$$z' = \mu' + z\sigma'$$

Infatti, il valore di aspettazione di z' risulta essere:

$$E[z'] = \mu' + \sigma' \cdot E[z] = \mu' + 0 = \mu'$$

La varianza invece:

$$\sigma_{z'}^2 = (\sigma')^2$$

Per cui, $f(z') = N(\mu', \sigma')$

Anche questo metodo presenta una difficoltà. Infatti, dal momento che la gaussiana è illimitata lungo l'asse delle ascisse, è necessario troncarla numericamente quando la si manipola al computer.

- Trasformata/metodo di Box-Mueller (per N(0,1) e Cauchy):

  1. Parto da due numeri r_1, r_2 generati attraverso una distribuzione uniforme nell'intervallo $[0, 1]$.
  2. Definisco:
  $$x_1 = \sqrt{-2\ln(r_1)}\cos(2\pi r_2)$$
  $$x_2 = \sqrt{-2\ln(r_1)}\sin(2\pi r_2)$$
  entrambi risultano seguire una distribuzione normale standard e sono completamente scorrelati l'uno dall'altro. Dunque, con questo metodo è possibile ottenere due numeri che seguono una distribuzione normale standard a partire da due che, invece, seguono una distribuzione uniforme.
  Il rapporto $\frac{x_1}{x_2}$ segue una distribuzione di Cauchy.

- Generatore lineare congruenziale: metodo di generazione di numeri pseudocasuali avente un certo periodo. Cioè, da un certo punto in poi, i numeri si ripetono. Per mantenere un certo grado di casualità, si cerca di prendere un periodo lunghissimo.

  Chi vuole sapere che cosa fa l'assistente nella vita, cercare "Deep Inelastic Scattering"

9 Note sulle esercitazioni

9.1 Esercitazione 5:

Polarizzazione = direzione privilegiata per lo spin.

Diffusione elastica protone-carbonio = quando un fascio collimato di protoni colpisce in modo elastico un bersaglio di carbonio, diffondendosi di conseguenza con un certo angolo (di diffusione).

Avremo una tabella di valori ottenuti da esperimenti reali, con un valore fisso di θ_1, tanti di θ_2 e di ε_p. Bisogna stimare i valori di a_c e a_p con il MML, considerando le variabili indipendenti, e le regioni corrispondenti a 1,2 e 3 deviazioni standard.

ε_p e ε_c sono delle asimmetrie misurate, chiamate anche "poteri analizzanti", sono legate all'ampiezza degli angoli azimutali, in funzione dei quali si possono calcolare. Le relazioni importanti sono:

$$\varepsilon_p = PA_p = a_c a_p \theta_1 \theta_{2p}$$
$$\varepsilon_c = PA_c = a_c^2 \theta_1 \theta_{2c}$$

dove:

$$A_p = a_p \theta_{2p} \qquad A_c = a_c \theta_{2c}$$

Possiamo immaginare le due asimmetrie come due rette con intercetta nulla. Il nostro obiettivo è estrarre i valori dei parametri a_c e a_p dalle espressioni scritte, perché vogliamo essere in grado di misurare la polarizzazione degli anti-protoni. Per stimare i parametri si usa il MML. La prima cosa da fare è scrivere la fz. di Likelihood.

Abbiamo delle misure degli angoli di diffusione e delle due asimmetrie con delle incertezze statistiche associate. Si assume che le misure delle asimmetrie siano indipendenti tra loro e che siano distribuite secondo una certa funzione di distribuzione di Gauss. Intanto, il Likelihood sarà il prodotto di molte gaussiane dato che le variabili casuali sono indipendenti:

$$L = \prod_{i=1}^{n_p} \frac{1}{\sqrt{2\pi\sigma_{p,i}^2}} \cdot \exp\left(-\frac{(\varepsilon_{p,i} - \varepsilon_p)^2}{2\sigma_{p,i}^2}\right) \cdot \prod_{j=1}^{n_c} \frac{1}{\sqrt{2\pi\sigma_{c,i}^2}} \cdot \exp\left(-\frac{(\varepsilon_{c,i} - \varepsilon_c)^2}{2\sigma_{c,i}^2}\right)$$

ed esprimiamo ε_p e ε_c usando le formule scritte prima:

$$L = \prod_{i=1}^{n_p} \frac{1}{\sqrt{2\pi\sigma_{p,i}^2}} \cdot \exp\left(-\frac{(\varepsilon_{p,i} - \mu_{p;i,j})^2}{2\sigma_{p,i}^2}\right) \cdot \prod_{j=1}^{n_c} \frac{1}{\sqrt{2\pi\sigma_{c,i}^2}} \cdot \exp\left(-\frac{(\varepsilon_{c,i} - \mu_{c;i,j})^2}{2\sigma_{c,i}^2}\right)$$

dove

$$\mu_{p;i,j} = a_c\theta_1 a_p\theta_{2p,i}$$
$$\mu_{c;i,j} = ac^2\theta_1\theta_{2c,i}$$

Tuttavia, questa è un'espressione difficile da manipolare. Passiamo al logaritmo in base naturale:

$$\ln(L) = \sum_{i=1}^{n_p} \ln\left[\frac{1}{\sqrt{2\pi\sigma_{p,i}^2}} \cdot \exp\left(-\frac{(\varepsilon_{p,i} - \mu_{p;i,j})^2}{2\sigma_{p,i}^2}\right)\right] + \sum_{j=1}^{n_c} \ln\left[\frac{1}{\sqrt{2\pi\sigma_{c,i}^2}} \cdot \exp\left(-\frac{(\varepsilon_{c,i} - \mu_{c;i,j})^2}{2\sigma_{,i}^2}\right)\right]$$

che diventa:

$$\sum_{i=1}^{n_p}\left(-\frac{1}{2}\ln(2\pi\sigma_{p,i}^2) - \frac{(\varepsilon_{p,i} - \mu_{p;i,j})^2}{2\sigma_{p,i}^2}\right) + \sum_{j=1}^{n_c}\left(-\frac{1}{2}\ln(2\pi\sigma_{c,i}^2) - \frac{(\varepsilon_{c,i} - \mu_{c;i,j})^2}{2\sigma_{c,i}^2}\right)$$

Cerchiamo di massimizzare questa funzione. Tuttavia la soluzione analitica non è né semplice né è garantita che esista. Visto che disponiamo di un computer, facciamo una griglia, sostituiamo manualmente i valori e vediamo dove la funzione assume il massimo valore. Come si sceglie la griglia? Sull'asse orizzontale si mettono i valori di a_p mentre su quello verticale i valori di a_c.

Sappiamo che:

$$\varepsilon_c = a_c^2 \cdot \theta_1 \cdot \theta_{2c}$$

Vediamo che possiamo scriverlo come:

$$y = n \cdot x$$

Quindi, possiamo applicare il metodo dei minimi quadrati per trovare una prima stima di a_c. Stesso discorso per a_p:

$$\varepsilon_p = a_p \cdot a_c \cdot \theta_1 \cdot \theta_{2p}$$

Si stima il prodotto dei primi tre termini, si sostituisce la stima di a_c e si trova la stima di a_p con la sua varianza. Poi, si suppone di costruire un rettangolo partendo dalle due stime, spostandosi di quattro deviazioni standard a sinistra e a destra per $\hat{a_p}$ e in alto e in basso per $\hat{a_c}$. Oppure, si costruisce in modo approssimativo la retta guardando come si dispongono i punti sperimentali con le loro incertezze statistiche, partendo dal grafico $\theta_{2c} - \varepsilon_c$:

$$y = m \cdot x \to \varepsilon_c(\theta_{2c} = 0.22) \approx 0.01 = a_c^2 \cdot 0.240 \cdot 0.22 \to a_c \approx \sqrt{\frac{0.01}{0.240 \cdot 0.22}} \approx 0.6$$

Si ripete la procedura per diversi altri punti appartenenti sperimentali per avere un'idea di come varia a_c. Per esempio:

$$\varepsilon_c(\theta_{2c} = 0.268) \approx 0.008 \to a_c \approx \sqrt{\frac{0.008}{0.240 \cdot 0.268}} \approx 0.4$$

Quindi si può considerare 0.6 ± 0.2. Lo stesso ragionamento si può fare con a_p, dal grafico $\theta_{2p} - \varepsilon_p$. La griglia viene, dunque, centrata nella coppia di valori di a_c e a_p.

Come si costruisce la griglia dal punto di vista pratico? Bisogna fare una certo numero di suddivisioni. Ma come si sceglie il passo della griglia? Non deve essere minore dell'incertezza statistica, quindi va bene 0.01. Per ogni punto della griglia va calcolato il logaritmo della funzione di Likelihood. Poi si trova il valore massimo, in corrispondenza del quale si trovano anche le stime dei parametri a_c e a_p.

Per determinare gli intervalli di confidenza, bisogna visualizzare tutti i valori dei parametri per cui:

$$ln(L) \approx ln(L_{max}) - \frac{1}{2}$$

o meglio:

$$ln(L) - (ln(L_{max}) - 1/2) \leq \varepsilon$$

dove ε è un numero arbitrariamente piccolo da scegliere manualmente. Infine, bisogna disegnare i grafici dei punti sperimentali con le incertezze associate.

Seconda parte dell'esercizio: Per determinare la probabilità che i valori dei parametri stiano proprio nelle regioni individuate, si usa il metodo delle repliche. Si prendono i punti sperimentali e da essi se ne costruiscono altri che si discostano dai primi di poco (vedi esercitazione 2). Quindi, si costruiscono tante repliche della tabella di dati e per ciascuna di essa si stimano i parametri a_c e a_p.

Nell'esercitazione 2, il grafico delle coppie $(\hat{m}, \hat{q})$ era un'ellisse, ma in questo caso non sappiamo quale forma abbia. Dunque, si costruisce una griglia un po' più grande che conti quante coppie cascano in ciascun quadratino. Si ottiene così un istogramma bidimensionale.

Quindi, si cerca il quadratino dove si ha il massimo dell'istogramma. Poi, si parte questo massimo e si scende di varie percentuali. Per esempio, si sommano i conteggi dei quadratini in cui si ottengono valori uguali o maggiori del minimo meno una certa percentuale. Si scende fino a quando non si trova 0.38 come probabilità sui conteggi totali.

.
.
.

L'obiettivo dell'esercitazione è misurare la polarizzazione degli antiprotoni.

$$p = a_c \cdot \theta_1$$

Si prendono tutti gli a_c stimati dalle repliche e si calcola per ciascuno la polarizzazione. Il grafico di quest'ultima avrà un andamento gaussiano, centrato nel valore di aspettazione:

$$\bar{p} = \bar{a_c} \cdot \theta_1$$

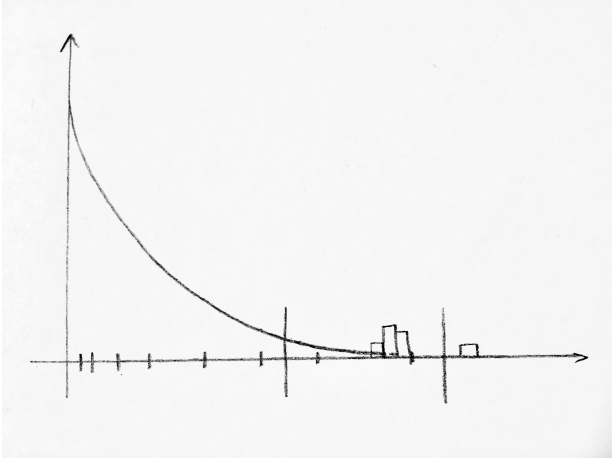
Vogliamo trovare l'intervallo di confidenza per la polarizzazione. Si devono sommare i conteggi dei bin estremi, dove "estremo" è stabilito in modo arbitrario. Se si trova 0.19 da una parte e dall'altra, allora va bene, altrimenti bisogna allargare o restringere a seconda dei casi. Una volta che all'interno si ha 0.68, si leggono i valori della polarizzazione in corrispondenza della suddivisione e si individua il valore di una deviazione standard.

Se aggiungendo due bin alla volta si trova una percentuale molto superiore al 68 per cento, si ripete tutto con dei bin meno larghi, quindi un'analisi più fitta.

9.2 Esercitazione 6:

Bisogna analizzare i dati relativi alle misurazioni dei raggi cosmici, forniti dai colleghi di fisica nucleare e subnucleare. In particolare, è necessario applicare metodo dei minimi quadrati o quello del *Maximum Likelihood*. Infine, si effettua il test di ipotesi. La seconda parte chiederà di usare la funzione di Erlang riferita a uno dei tre eventi, stima di nuovo di parametro.

Nell'esercitazione 6 ci si aspetta, appunto, una distribuzione di Erlang (distribuzione diminuisce all'aumentare del tempo), quindi la probabilità di avere degli eventi sulle code è molto bassa, quindi facendo un istogramma si troverebbero molte colonne vuote in corrispondenza di quella zona. Per fare il test di ipotesi è meglio partire da zero con delle colonnine strette e man mano allargarle, invece di averle tutte della medesima larghezza.



10 Tabelle

	Gaussiana	Esponenziale	Binomiale	Poisson	χ^2_n
Distribuzione	$\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}$	$\frac{1}{\tau}e^{-x/\tau}$	$\binom{n}{k}p^k q^{n-k}$	$\frac{\nu^k}{k!}e^{-\nu}$	$\frac{x^{\frac{n}{2}-1}e^{-x/2}}{2^{n/2}\Gamma\left(\frac{n}{2}\right)}$
Generatrice μ_k	$e^{\sigma^2 t^2/2}$	$\frac{e^{-\tau t}}{1-t\tau}$	$(pe^{tq}+qe^{-tp})^n$	$e^{\nu(e^t-1-t)}$	$e^{-tn}(1-2t)^{-n/2}$
Generatrice μ_k^*	$e^{\frac{\sigma^2 t^2}{2}+\mu t}$	$\frac{1}{1-t\tau}$	$(pe^t+q)^n$	$e^{\nu(e^t-1)}$	$(1-2t)^{-n/2}$
μ_0	1	1	1	1	1
μ_1	0	0	0	0	0
μ_2	σ^2	τ^2	npq	ν	$2n$
μ_3	0		$npq(q-p)$	ν	$8n$
μ_4	$3\sigma^2$		$\frac{1-6pq}{npq}$	$3\nu^2+\nu$	$12n^2+48n$
γ_1	0	2	$\frac{q-p}{\sqrt{npq}}$	$\frac{1}{\sqrt{\nu}}$	$\frac{8}{2\sqrt{2n}}$
γ_2	0	6		$\frac{1}{\nu}$	$\frac{12}{n}$
μ_0^*			1		
$\mu_1^* = \mu_x$	μ	τ	np	ν	
μ_2^*					
μ_3^*					
μ_4^*					

Tabella 8: Funzioni di distribuzione e numeri utili