Laboratory 2

December 23, 2021

Contents

Lecture 1: Introduzione

gio 07 ott 2021 13:30

1 Introduzione

La statistica utilizza la probabilità, ma non ha a disposizione tutte le informazioni: prende un campione ed inferisce delle informazioni.

Definizione. Un evento casuale è il risultato di un esperimento o di un'osservazione che non può essere previsto con certezza.

Esso è la base della probabilità in quanto:

- dev'essere ripetibile (cioè osservare l'avvenimento più volte);
- deve presentare delle modalità mutualmente esclusive;
- deve essere singolarmente non prevedibile.

Definizione. Lo spazio campionario (sample place, spazio delle fasi in alcuni contesti) è un insieme Ω di tutte le possibili modalità con cui l'evento può accadere.

Definizione. La popolazione è l'insieme di tutti i possibili eventi.

Definizione. Un campione è l'insieme degli eventi casuali raccolti.

Gli insiemi degli eventi si possono rappresentare tramite i diagrammi di Eulero-Venn.i

Definizione. assiomatica di probabilità. Si dà una definizione assiomatica della probabilità attraverso gli assiomi di Kolmogorov La probabilità è una funzione $P : \Omega \to [0, 1]$ che rispetta le seguenti proprietà:

- $P(A) \geq 0$, $\forall A \subset \Omega$;
- $P(\Omega) = 1$;
- $P(A \cup B) = P(A) + P(B)$ quando $A \cap B = \emptyset$.

Conseguenze.

- $P(A) = 1 - P(A^c)$:
$$1 = P(\Omega) = P(A \cup A^c = P(A) + P(A^c).$$
- $P(A) \leq 1$:
$$P(A^c) \geq 0 \implies P(A) = 1 - P(A^c) \leq 1.$$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

dimostra.

Definizione. classica di probabilità. La probabilità di un evento A è definita come il rapporto tra i casi favorevoli ed i casi totali.

Definizione. frequentista di probabilità. In un campione di N eventi, la probabilità associata all'evento di tipo A è la frazione di casi per cui A avviene calcolata per $N \to \infty$:

$$P(A) = \lim_{N \to \infty} \frac{n(A)}{N}.$$

La definizione di limite in statistica è diversa da quella dell'analisi matematica.

Osservazioni.

- Per $N \to \infty$ il campione tende alla popolazione.
- La probabilità così definita soddisfa gli assiomi di Kolmogorov.
- Se si parte dalla definizione classica allora la definizione frequentista è il teorema di Bernoulli o legge dei grandi numeri.

[rivedi] La definizione frequentista di probabilità è la statistica frequentista, ma è limitante: spesso si desidera associare una probabilità non all'accadere di un evento, ma alla veridicità di una affermazione (come "domani piove").
L'altro approccio è quello Bayesiano. [rivedi] Tuttavia, nella formulazione della statistica come estensione della logica matematica, si introduce la plausibilità di una teoria. È possibile introdurre in modo rigoroso l'utilizzo delle conoscenze a priori che is hanno riguardo il fenomeno che si sta studiando.

1.1 Probabilità condizionata

Si consideri una funzione di probabilità $P : \Omega \to [0, 1]$. Si A, B due eventi in Ω non disgiunti. Allora la probabilità condizionata è

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

Si normalizza la probabilità dell'intersezione sulla probabilità di A: si riduce il sample space.
Inoltre

- $P(X)$ possiede X come variabile indipendente.
- $P(X|A)$ è una nuova funzione con X variabile indipendente e A come parametro.
- $P(\dots|A)$ soddisfa gli assiomi.

Proposizione. Due eventi sono indipendenti se e solo se $P(B|A)$ è la stessa per ogni A: $P(B|A) = P(B|\Omega) = P(B)$.
Due eventi sono indipendenti se e solo se $P(A \cap B) = P(A) \cdot P(B)$.

Teorema di Bayes. Si consideri

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

Dato che $A \cap B = B \cap A$ risulta

$$P(A \cap B) = P(A)P(B|A) = P(B \cap A) = P(B)P(A|B) \implies P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Il teorema di Bayes applicato alla fisica assume la forma

$$P(\text{teoria}|\text{risultato}) = \frac{P(\text{risultato}|\text{teoria})}{P(\text{risultato})} P(\text{teoria}).$$

1.2 Variabili aleatore discrete e continue

In fisica, l'evento casuale è identificato con una random variable. Essa può essere discreta oppure continua.

Random variable discreta. Si ha

- $\Omega \subset \mathbb{N}$
- $P(k)$, probabilità dell'evento k
- $P : \Omega \subset \mathbb{N} \to [0, 1]$, $P(k) > 0$, $\forall k$

[rivedi]

Random variable continua. Si ha

- $\Omega \subset \mathbb{R}$
- $P(a < x < b)$, probabilità di trovarsi entro un certo intervallo
- Si definisce una densità di probabilità, pdf, tale per cui $P(a < x < b) = \int pdf(x)\,\mathrm{d}x$

Lecture 2

gio 14 ott 2021 13:30

1.3 Distribuzioni di probabilità

Se l'evento casuale è descritto con un numero reale x si ha una random variable per cui vale

- $\Omega \subset \mathbb{R}$;
- $P(x)$ è un infinitesimo in quanto Ω ha una quantità infinita di eventi;
- La probabilità associata ad un intervallo è $P(a < x < b) \geq 0$.

Esempio. La temperatura T è un numero reale espresso da un numero limitato di cifre significative dettato dalla sensibilità dello strumento.
Ripetendo N volte le misure si ha:

- I valori sono registrati attorno ad un valor medio (questo vale supponendo che le fluttuazioni siano confrontabili con la sensibilità dello strumento).
- Si hanno delle classi di frequenza: si conta il numero di misure in un certo intervallo, $f = \frac{n}{N}$.
- Si ottiene un istogramma che ha area unitaria per costruzione.

Nell'istogramma, l'area di un bin è la probabilità che una misura cada in tale bin. La lunghezza di un intervallo può essere ridotta (magari aumentando di pari passo la sensibilità dello strumento). Aumentando la quantità di misure e diminuendo la larghezza dei bin si ottiene una curva continua la cui area è unitaria: si ottiene una funzione di densità di probabilità (*pdf*, probability density function).

Per una random variable continua si definisce *pdf* una funzione:

$$pdf : \Omega \subset \mathbb{R} \to \mathbb{R}^+.$$

Mediante la quale si calcola la probabilità:

- $P(x < RV < x + \mathrm{d}x) = pdf \cdot \mathrm{d}x$
- $P(a < x < b) = \int_a^b pdf(x) \cdot \mathrm{d}x \leq 1$

Per essere una *pdf* bisogna soddisfare gli assiomi di Kolmogorov:

- $pdf(x) > 0$, $\forall x$;
- $\int_\Omega pdf(x)\,\mathrm{d}x = 1$;

- $P(a < x < b) = \int_a^b pdf(x)\,\mathrm{d}x$.

Si definisce la funzione cumulativa di densità (cdf, cumulative density function) come

$$cdf(x) = \int_a^x pdf(x)\,\mathrm{d}x.$$

Di solito $a \to -\infty$. La *cdf* è la primitiva della *pdf* e restituisce una probabilità

$$pdf(x) = \frac{\mathrm{d}}{\mathrm{d}x} cdf(x) \iff cdf(x) = \int_a^x pdf(x)\,\mathrm{d}x.$$

Alcuni parametri caratteristici sono:

- Valor medio,
- moda,
- mediana,
- varianza,
- skewness (obliquità),
- kurtosis (gobba).

Conoscere solo l'espressione di una *pdf* non è significativo. Quando si costruiscono modelli complicati è spesso impossibile trovare una forma analitica; dunque si utilizzano tali parametri.

Sia $u(x)$ una funzione con random variable indipendente x. Il valore atteso di $u(x)$ è

$$E[u(x)] = \int_\Omega u(x)\, pdf(x)\,\mathrm{d}x.$$

Essa è una media pesata sulla *pdf*. L'operatore del valor atteso è lineare:

- $E[u(x) + v(x)] = E[u(x)] + E[v(x)]$;
- $E[ku(x)] = kE[u(x)]$.

Media. La media μ è il valore atteso di x :

$$E[x] = \int_\Omega x\, pdf(x)\,\mathrm{d}x = \mu.$$

Essa equivale alla media aritmetica di x ottenuta considerando l'intera popolazione. La media non va confusa con la media campionaria.

Varianza. La varianza σ^2 è

$$E[(x-\mu)^2] = \int_\Omega (x-\mu)^2\, pdf(x)\,\mathrm{d}x = \sigma^2.$$

La varianza è anche detta scarto quadratico medio. La deviazione standard σ è legata alla larghezza della *pdf*. Inoltre vale

$$\sigma^2 = E[x^2] - E[x]^2.$$

Moda e mediana. La moda è il massimo della *pdf*. Mentre la mediana divide a metà la *pdf*. La moda può avere più di un valore: si parla di distribuzioni multimodali (bi-, tri-, ...).

La media è adatta a manipolazioni matematiche; tuttavia, essa è influenzata da valori estremi. La mediana è l'opposto.

Momento. Il momento di ordine m è

$$E[x^m] = \int_\Omega x^m \, pdf(x) \, \mathrm{d}x.$$

Il momento di ordine 1 è la media. I momenti centrali sono

$$E[(x-\mu)^m] = \int_\Omega (x-\mu)^m \, pdf(x) \, \mathrm{d}x.$$

Il momento centrale di ordine

- 1 è sempre nullo;
- 2 è la varianza, legata alla larghezza;
- 3 è la skewness $\gamma_1 = \frac{E[(x-\mu)^3]}{\sigma^3}$, legata all'asimmetria;
- 4 è la kurtosis $\gamma_2 = \frac{E[(x-\mu)^4]}{\sigma^4} - 3$, legata al picco e si prende come riferimento la gaussiana.

Se esistono i momenti centrali allora questi identificano univocamente la *pdf* e ciò è importante quando la *pdf* non è analitica, ma si ha una stima dei momenti centrali.

Riproduttività. Si considerino x, y random variable con medesima *pdf*, ma diversi parametri. Se $z = x + y$ ha la stessa *pdf* allora la *pdf* gode della proprietà di riproduttività.

1.3.1 Distribuzioni notevoli

La modellizzazione di specifici processi aleatori porta a costruire specifiche *pdf* che li descrivono. La $pdf(x; \alpha, \ldots)$ è caratterizzata da

- una forma funzionale;
- parametri, spesso espressi in termini dei momenti.

Uniforme. La distribuzione uniforme descrive una random variable con stessa densità di probabilità per ogni elemento di Ω. I PRNGs (pseudo-random number generators) hanno distribuzioni quasi uniformi. Si vede successivamente come a partire da una distribuzione uniforme si può ricavare una distribuzione arbitraria.
Sia $\Omega = [a, b]$, $pdf(x) = k$, $\forall x \in \Omega$. Dalla condizione di normalizzazione segue

$$\int_\Omega pdf(x) \, \mathrm{d}x = \int_a^b k \, \mathrm{d}x = 1 \iff k = \frac{1}{b-a}.$$

Dunque si dice distribuzione uniforme

$$U(x; a, b) = \frac{1}{b-a}.$$

I parametri sono gli estremi dell'intervallo. Inoltre

$$E[x] = \frac{a+b}{2}, \quad Var[x] = \frac{(b-a)^2}{12}.$$

Gode della proprietà di riproduttività e

$$cdf(x) = \int_a^x \frac{\mathrm{d}x}{b-a} = \frac{1}{b-a}(x-a).$$

Esempio. Si supponga di misurare una temperatura con un termometro con una bassa sensibilità. Le fluttuazioni del campione sono molto piccole. Dunque, il risultato ripetutamente misurato è sempre lo stesso. La misura di x_0 è di sicuro tra $x_0 - 0.5$ e $x_0 + 0.5$. Ma questa considerazione è diversa dall'incertezza casuale perché la distribuzione è uniforme. Il valore reale è qualsiasi tra $[x_0 - 0.5, x_0 + 0.5]$. Dunque, se si volesse confrontare tale valore con uno casuale bisogna usare la deviazione standard. Inoltre, la probabilità entro $1\,\sigma$ non è sempre 68%, ma è comunque prossima a tale valore.

Gaussiana. Se le barre di errore fissano la sensibilità allora un fit deve passare per tutte le barre di errore perché si ha la certezza che la misura sia all'interno dell'intervallo di incertezza. Per errori statistici è insolito che il fit passi per tutte le barre; qualora sia il caso, si potrebbe aver sovrastimato l'errore.
Inoltre

$$E[x] = \mu, \quad Var[x] = \sigma, \quad \gamma_1 = 0, \quad \gamma_2 = 0.$$

Per la distribuzione gaussiana si ha un valore di Full Width at Half-Maximum (FWHW) di $2\sqrt{2\ln 2}$. Essa è anche riproduttiva.
Si definisce la distribuzione gaussiana standardizzata: $\mu = 0$, $\sigma = 1$. Una qualsiasi gaussiana può essere riportata ad una gaussiana standard.
Inoltre

$$cdf(x) = \int_{-\infty}^{x} G(x)\,\mathrm{d}x = \mathrm{erf}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x^2}{2}}.$$

Breit-Wigner o lorentziana. Si ha

$$BW(x; \alpha, M) = \frac{1}{\pi\alpha} \frac{1}{1 + \frac{(x-M)^2}{\alpha^2}}.$$

Non ha media né momenti. Essa è la distribuzione da considerare per particelle ed è simmetrica rispetto ad M.

Lecture 3

gio 21 ott 2021 13:30

2 Teorema centrale del limite e trasformazioni di variabili

Campionamento.
Si consideri una popolazione di eventi che si possono classificare in un sample space. Campionare significa estrarre casualmente dalla popolazione così da costituire un campione. Molte volte la popolazione è infinita, quindi è utile utilizzare una distribuzione. Bisogna trovare un algoritmo che campioni rispettando la distribuzione della popolazione. Si definisce il campionamento di una *pdf* dove la random variable x è un numero reale.

2.1 Teorema centrale del limite.

La funzione di una random variable è anch'essa una random variable. Il teorema entrale del limite studia la *pdf* di una random variable funzione di random variables. Si considerino N distribuzioni con μ, σ^2 finite; si estragga un campione da ciascuna, allora

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

è una random variable. In particolare, $\overline{x}$ è distribuita in modo gaussiano per $N \to +\infty$ con media pari alla somma delle medie e varianza pari alla somma delle varianze.

La media e la varianza di $\overline{x}$ si possono calcolare senza conoscere la *pdf* e anche per N finito (si considera che medie e varianze siano indipendenti):

- Media $\mathbb{E}[\overline{x}] = \mu_{\overline{x}} = \frac{\sum \mu_i}{N}$ detta valor di aspettazione.
- Varianza $\mathrm{Var}[\overline{x}] = \sigma^2_{\overline{x}} = \frac{\sum \sigma_i^2}{N^2}$.

Per dimostrare tale fatto si utilizza la linearità dell'operatore di aspettazione quando gli eventi sono indipendenti $\mathbb{E}[x_1 + x_2] = \mathbb{E}[x_1] + \mathbb{E}[x_2]$. Infatti

$$\begin{aligned}\mathbb{E}[x_1 + x_2] &= \int (x_1 + x_2)\, pdf(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 \\ &= \int x_1\, pdf_1\, \mathrm{d}x_1 \int pdf_2\, \mathrm{d}x_2 + \int x_2\, pdf_2\, \mathrm{d}x_2 \int pdf_1\, \mathrm{d}x_1 = \mathbb{E}[x_1] \cdot 1 + \mathbb{E}[x_2] \cdot 1.\end{aligned}$$

Dunque

$$\mathbb{E}[\overline{x}] = \mathbb{E}\left[\frac{\sum x_i}{N}\right] = \frac{\sum \mathbb{E}[x_i]}{N} = \frac{\sum \mu_i}{N}.$$

ed in modo analogo per la varianza. Per il TCL, per $N \to +\infty$ si ottiene una gaussiana

$$pdf(\overline{x}) = pdf\left(\frac{\sum x_i}{N}\right) \to G(\mu_{\overline{x}}, \sigma_{\overline{x}}).$$

Bisogna notare che il TCL non afferma che la somma di N distribuzioni tenda ad una gaussiana.

Applicazioni del TCL. La grandezza fisica è affetta d un errore casuale. In altri casi, la lettura è sempre la stessa, si ha comunque l'errore dovuto ala sensibilità. Si consideri x_0 il vero valore; x il risultato di una misura; ε l'errore associato a tale misura che è la somma di tanti ε_i errori casuali con *pdf* e dunque ε è una variabile aleatoria. La variabile $x = x_0 + \varepsilon$ è aleatoria con una *pdf* associata.
Ripetere la misura significa campionare la $pdf(x)$ che dipende $pdf(\varepsilon)$. Se ε è la somma di errori casuali, ciascuno con una *pdf*, allora il TCL afferma che $pdf(\varepsilon)$ è gaussiana. La $pdf(\varepsilon)$ è centrata in zero, altrimenti si ha un errore sistematico. La media della $pdf(x)$ è il valore vero.

Un'altra applicazione del TCL è la generazione di numeri casuali. Tramite una *pdf* uniforme si può passare ad una *pdf* gaussiana sommando i valori campionati dalla *pdf* di partenza.

Da una random variable x descritta da una $pdf_1(x)$ si vuole ricavare $pdf_2(y)$ associata ad una random variable ottenuta mediante la trasformazione $x \mapsto y(x)$. Possono sorgere due problemi:

- calcolare $pdf_2(y)$ analiticamente;
- determinare μ_y e σ_y, cioè propagare gli errori.

Esempio. Dalla $pdf(x)$ uniforme si vuole passare ad una $pdf(y)$ tale che $y = \cos x$ sull'intervallo $[0, \frac{\pi}{2}]$. Si campiona $pdf(x)$ e si osserva $y = \cos x$. I punti di $pdf(y)$ si aggregano vicino a 1 quindi l'asse y non è distribuito uniformemente.

2.2 Cambio di variabile.

Si studia il cambio di variabile in una dimensione. Sia x una random variable descritta da una pdf_1. Sia $y = y(x)$ una funzione biunivoca (quindi monotona) e derivabile. Allora y è una random variable distribuita con $pdf_2(y)$ data da

$$pdf_2(y) = |x'(x)| pdf_1(x).$$

Dimostrazione. Si consideri un intervallo infinitesimo $\mathrm{d}x$ e sia $\mathrm{d}y$ l'intervallo trasformato. Ai due intervalli si associa la stessa probabilità:

$$pdf_1(x)\, \mathrm{d}x = pdf_2(y)\, \mathrm{d}y \iff pdf_2(y) = \left|\frac{\mathrm{d}x(y)}{\mathrm{d}y}\right| pdf_1(x) = |x'(y)| pdf_1(x) = \left|\frac{1}{y'(x)}\right| pdf_1(x).$$

Esempio. Sia x uniforme in $[0, \pi]$, con $pdf = \frac{1}{\pi}$; sia $y = \cos x$. Allora

$$pdf(y) = \frac{1}{x}\left|\frac{\mathrm{d}}{\mathrm{d}y}\operatorname{acos} y\right| = \frac{1}{\pi}\frac{1}{\sqrt{1-y^2}}.$$

Esempio. Si vede la distribuzione normale. Sia $y = \frac{x-\mu}{\sigma}$, $x = \sigma y + \mu$, $x' = \sigma$ allora:

$$G(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \implies pdf(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}.$$

Esempio. Si vede la distribuzione log-normale. Sia x un random variable con distribuzione gaussiana. Sia $y = e^x$, $x = \ln y$, $x' = \frac{1}{y}$. Allora

$$pdf(y) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{y} e^{-\frac{(\ln y-\mu)^2}{2\sigma^2}}.$$

Con momenti

$$\mathbb{E}[y] = e^{\mu+\frac{\sigma^2}{2}}, \quad \mathrm{Var}[y] = e^{2\mu+\sigma^2}(e^{\sigma^2}-1).$$

Nel caso in cui si voglia ottenere solamente μ_y e σ_y^2 si propagano gli errori.

Trasformazioni lineari. Sia $v = \frac{L}{t}$ la velocità con il tempo t misura e la lunghezza L fissa. Sia $y(x)$ una funzione lineare in x, $y = ax + b$, si usa la linearità dell'operatore di aspettazione:

$$\mu_y = \mathbb{E}[y] = \int y\, pdf(y)\,\mathrm{d}y = \int y(x)\, pdf(x)\,\mathrm{d}x = \int ax+b\, pdf(x)\,\mathrm{d}x = a\mathbb{E}[x]+b = a\mu_x+b = y(\mu_x).$$

Similmente $\sigma_y^2 = a^2\sigma_x^2$.

Trasformazioni non lineari. Per mezzo di Taylor risulta

$$y(x) = y(\mu_x) + \frac{\mathrm{d}y(\mu_x)}{\mathrm{d}x}(x-\mu_x) + o(x).$$

Dunque la media risulta essere

$$\begin{aligned}\mu_y = \mathbb{E}[y] = \int y\, pdf(y)\,\mathrm{d}y = \int y(x)\, pdf(x)\,\mathrm{d}x \approx \int y(\mu_x)\, pdf(x) + \frac{\mathrm{d}y(\mu_x)}{\mathrm{d}x}(x-\mu_x)\, pdf(x)\,\mathrm{d}x \\ = y(\mu_x) + \frac{\mathrm{d}y(\mu_x)}{\mathrm{d}x}\int (x-\mu_x)\, pdf(x)\,\mathrm{d}x = y(\mu_x).\end{aligned}$$

L'ultimo integrale è nullo perché esso è il momento centrale del primo ordine.
Tuttavia, al secondo ordine i termini non si annullano:

$$\begin{aligned}\mu_y &\approx y(\mu_x) + \frac{1}{2}\frac{\mathrm{d}^2 y(x)}{\mathrm{d}x^2}\sigma_x^2 \\ \sigma_y^2 = \mathbb{E}[y^2] - \mu_y^2 &\approx \left(\frac{\mathrm{d}y(\mu_x)}{\mathrm{d}x}\right)^2 \sigma_x^2.\end{aligned}$$

Tuttavia, la *pdf* dev'essere piccata nella regione in cui si ha lo sviluppo di Taylor perché altrimenti l'approssimazione non risulta essere buona.

2.3 Random variables discrete

Si definisce discreta una random variabile per cui l'evento casuale è identificabile con un numero intero k:

- il sample space $\Omega \subset \mathbb{N}$;
- $P(k)$ è la probabilità che la random variable assuma il valore k;
- $P : (\Omega \subset \mathbb{N}) \to [0, 1]$.

Gli assiomi di Kolmogorov diventano

- $P(k) > 0$, $\forall k$;
- $\sum P(k) = 1$;
- $P(k \vee h) = P(k) + P(h)$, per $h, k \in \Omega$, $h \neq k$.

Bernoulli trials. Il caso più semplice di probabilità discreta è quello in cui sono possibili solo due eventi (uno complementare dell'altro):

- Il successo ha probabilità p;
- l'insuccesso ha probabilità $q = 1 - p$;
- il sample space $\Omega = \{0, 1\}$.

Distribuzione binomiale. La distribuzione binomiale descrive la probabilità di k successi in N prove:

$$B(k; N, p) = \binom{N}{k} p^k (1-p)^{N-k}.$$

Le prove singole sono dette Bernoulli trials. Il sample space è un intervallo di interi finito $\Omega = [0, N] \subset \mathbb{N}$.

Dimostrazione. A grandi linee, le prove (Bernoulli trials) sono tutte indipendenti tra di loro e ciascuna ha la stessa probabilità di successo p e di insuccesso $(1p)$. L'espressione per la distribuzione binomiale si ricava contando la frazione di successi in tutti i casi possibili.

Per $N \to +\infty$, la distribuzione binomiale tende alla gaussiana. La media e la varianza sono

$$\mu = \sum_{k=0}^{N} k \cdot B = Np, \quad \sigma^2 = Np(1-p).$$

Mentre la skewness e la kurtosis sono

$$\gamma_1 = \frac{1-2p}{\sqrt{Np(1-p)}}, \quad \gamma_2 = \frac{1-6p(1-p)}{Np(1-p)}.$$

Inoltre, gode della proprietà di riproduttività e per $N \to +\infty$ si ha $\gamma_{1,2} \to 0$.

Lecture 4

gio 28 ott 2021 13:30

3 Poisson e *pdf* multidimensionali

La distribuzione binomiale descrive la probabilità di k successi in N prove

$$B(k, N, p) = \binom{N}{k} p^k (1-p)^{N-k}.$$

Dove k varia da zero fino ad N cioè il sample space è limitato. Il successo (numero 1) ha probabilità p e l'insuccesso (numero 0) ha probabilità $1-p$. Dunque k è la somma dei punteggi ottenuti in N prove.
I Bernoulli trials hanno due valori nel sample space e per ogni valore di k si associa una probabilità così che quella totale rappresenta la normalizzazione della probabilità.
Si estrae da una *pdf* che è dei Bernoulli trials e si sommano, così per il TCL si ottiene una distribuzione gaussiana. La distribuzione di k per N finito è la distribuzione binomiale. Per $N \to +\infty$ la binomiale è asintotica ad una Gaussiana perché k è la somma di tante variabili estratte dalla *pdf* dei Bernoulli trials che ha media e varianza finite, e le estrazioni sono indipendenti tra loro. La gaussiana ottenuta dalla binomiale $B(k; N, p)$ è $G(k; \mu, \sqrt{\mu})$, con $\mu = Np$; quindi fissato μ si conosce sia la media che la varianza della gaussiana. Inoltre k passa da essere una variabile discreta ad una variabile continua. Ma per $N \to +\infty$ la discretizzazione tende alla continuità.
È importante studiare l'asintoticità di una *pdf* perché è più facile lavorare sulla gaussiana che la binomiale perché è difficile gestire dei fattoriali con il calcolo numerico.
L'altra proprietà di asintoticità della binomiale è la Poissoniana. Per $p \to 0$ e Np finito, la binomiale è asintotica alla distribuzione di Poisson: $P(k; \lambda)$, con $\lambda = Np$. La dimostrazione di tale fatto è per induzione.

3.1 Distribuzione di Poisson

Essa descrive la probabilità di contare k eventi un intervallo Δx unitario nel caso di un processo per cui la frequenza media è costante e pari a λ e l'accadere di un evento è indipendente dall'accadere dell'evento in un istante successivo:

$$P(k; \lambda) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}.$$

Il parametro λ che è la media della *pdf* è definito come il numero medio di conteggi registrati in un $\Delta x = 1$

$$\lambda = \underline{E[k]}.$$

[rivedi]

La distribuzione di Poisson si applica ad una grande varietà di casi:

- raggi cosmici che attraversano un volume nell'unità di tempo
- bombe cadute per unità di superficie su Londra durante la seconda guerra mondiale
- errori tipografici
- persone con una malattia rara all'interno di un gruppo

Il modello su cui si basa la distribuzione di Poisson è

1 la probabilità che un evento avvenga in un intervallo $\mathrm{d}x$ è proporzionale all'ampiezza dell'intervallo $p\,\mathrm{d}x$; dove p è una costante indipendente da x (questo è quanto si intende dicendo che è un processo in cui la frequenza media è costante);

2 dalla proposizione precedente deriva anche che la probabilità di avere più di un evento nell'intervallo $\mathrm{d}x$ è nulla (per questo si parla spesso di eventi rari).

3 gli eventi sono indipendenti tra loro
il fatto che un evento sia avvenuto al momento x non influisce sulla probabilità che un nuovo evento avvenga o sia avvenuto prima o dopo il momento x.

Un modo per costruire la distribuzione di Poisson è il seguente

- ipotesi 1, la probabilità di 1 evento in $\mathrm{d}x$ è $p\,\mathrm{d}x$
- ipotesi 1 e 2, la probabilità di zero eventi in $\mathrm{d}x$ è $1 - p\,\mathrm{d}x$
- si definisce la probabilità di zero eventi in $[0, x]$ come $q(x)$
- dall'ipotesi 3, la probabilità di zero eventi in $[0, x + \mathrm{d}x]$ è $q(0, x + \mathrm{d}x)$ cioè il prodotto di due probabilità (eventi indipendenti):

$$q(x + \mathrm{d}x) = q(x) \cdot (1 - p\,\mathrm{d}x) \implies \frac{q(x + \mathrm{d}x) - q(x)}{\mathrm{d}x} = -pq(x).$$

Che ha soluzione $q(x) = q(0)e^{-px}$. Inoltre $q(0)$ si ricava imponendo che $q(x)$ sia una probabilità ed imponendo che $\int_\Omega q(x) = 1$. Dunque la distribuzione è perfettamente definita e continua. Questa è la distribuzione esponenziale che è strettamente legata alla distribuzione di Poisson. Per eventi che occorrono in modo Poissoniano, la probabilità di non avere eventi un intervallo $[0, x]$ è data da un esponenziale.
Si completa la dimostrazione che porta alla distribuzione di Poisson. Trovata l'espressione per $q(x)$ si passa a ricavare

- la probabilità di avere: 0 eventi in $[0, x_1]$ e 1 evento in $x_1, x_1 + \mathrm{d}x_1]$ che è $e^{-px_1}(p\,\mathrm{d}x_1)$
- la probabilità di avere k eventi in $[0, t]$ è $e^{-px}\frac{(px)^k}{k!}$.

Solitamente si sceglie di definire $\lambda = px$ ottenendo quindi l'espressione della Poissoniana.
Bisogna ricordarsi bene le ipotesi da quali si parte e da tutto questo discende che la probabilità di non avere eventi è una distribuzione esponenziale. La distribuzione della distanza tra due eventi Poissoniani è esponenziale.
La probabilità di non avere enti in un intervallo $[0, x]$ segue una distribuzione esponenziale. La probabilità di avere k eventi in un intervallo $[0, k]$ segue a distribuzione di Poisson.

Momenti.

- media λ
- varianza $\lambda = \mu$
- skewness $\gamma_1 = \frac{1}{\sqrt{\lambda}}$
- kurtosis $\gamma_2 = \frac{1}{\lambda}$
- riproduttività

La distribuzione di Poisson è asintotica alla distribuzione di Gauss per λ grande: $G(k; \lambda, \sqrt{\lambda})$, con $\mu = \lambda$ e $\sigma^2 = \lambda$. Si passa da una random variable intera ad una continua.

Esiste un comportamento asintotico che porta la binomiale a somigliare alla Poissoniana. Per la binomiale k è minore strettamente di N, per la binomiale k può anche essere maggiore.

3.2 Distribuzione esponenziale

La distribuzione esponenziale si scrive spesso nella forma

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

Essa misura la probabilità di non avere eventi in un intervallo $[0, t]$ quando la probabilità con cui un evento avviene in un intervallo $\mathrm{d}t$ è $p = \frac{1}{\tau}$. La media è τ e la varianza è τ^2.

Il decadimento di un nucleo è di per sé probabilistico. La vita media di un nucleo è il reciproco della probabilità. Tali distribuzioni sono utilizzate sia per fenomeni di meccanica quantistica, ma anche perché la binomiale descrive gli istogrammi.

3.3 *pdf* in n-dimensioni

Il sample space è multidimensionale. Una funzione generica che sia una *pdf* è scritta come $pdf(x)$. Se si cambia l'argomento, cambia anche la forma funzionale: $pdf(x) = f(x)$ e $pdf(y) = g(y)$ ma $f(z) \neq g(z)$.

Quando un evento è identificato da un vettore, inveche che da una varaibile monodimensionale, si parla di *pdf* multidimenisonali. Devono valere gli assiomi di Kolmogorov. Nel caso continuo:

$$pdf(\vec{x}) : \mathbb{R}^n \to \mathbb{R}^+.$$

La $cdf(\vec{x})$ ed i momenti sono definiti per estensione di quanto fatto nel caso monodimensionale:

$$cdf(\vec{x}) = \int_{-\infty}^{\vec{x}} pdf(\vec{x})\, \mathrm{d}\vec{x}.$$

La media della *pdf* è un vettore $\vec{\mu}$

$$\vec{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} E[\vec{x}] = \int \vec{x}\, pdf(\vec{x}) = \begin{pmatrix} E[x_1] \\ \vdots \\ E[x_n] \end{pmatrix}.$$

dove i è definito il valore di aspttazione per una singola componente x_i del vettore $\vec{\mu}$ come

$$E[x_i] = \int x_i\, pdf(\vec{x})\, \mathrm{d}\vec{x}.$$

La varianza è sostituita da una matrice $n \times n$ simmetrica, chiamata matrice di covarianza:

$$\sigma_{ij}^2 = E[(x_i - \mu_i)(x_j - \mu_j)].$$

I termini sulla diagonale sono l'equivalente della varianza.

Caso bidimensionale. Spesso, si parla di una *pdf* congiunta (joint *pdf*) per le random variables $\{x_1, \ldots, x_n\}$. L'evento è identificato da n variabili random monodimensionali. Nel seguito si considera il caso bidimensionale $pdf(x, y)$.

Probabilità congiunta. Nel caso bidimensionale la probabilità congiunta è una superficie in uno spazio tridimensionale:

$$pdf(x, y) : (\mathbb{R} \times \mathbb{R}) \to \mathbb{R}^+.$$

Probabilità marginale. La $pdf_M(x)$ è la *pdf* per x, indipendente dai valori assunti da y, è ottenuta integrando su una delle due variabili

$$pdf_{M_x}(x) = \int pdf(x, y)\,\mathrm{d}y.$$

Probabilità condizionata. La $pdf(x|y = y_0)$ è la *pdf* associata ad x quando y assume uno specifico valore y_0 (cioè è unzione della sola x):

$$pdf(x|y = y_0) = \frac{pdf(x, y_0)}{pdf_{M_y}(y_0)} = \frac{pdf(x, y_0)}{\int pdf(x, y_0)\,\mathrm{d}x}.$$

Si divide per la probabilità marginale così da normalizzare.

Relazione con la rappresentazione di Eulero. Le definizioni ora date sono solo un caso particolare di quelle discusse nella prima lezione quando si sono considerate

- una probabilità $P : \Omega \to [0, 1]$
- due eventi A e B in Ω non disgiunti

Si è parlato di

- $P(A \cap B)$ joint *pdf*
- $P(A)$ marginal *pdf*
- $P(A|B)$ conditional *pdf*

Si ipotizzi che gli eventi in Ω siano associa a due random variables differenti, x ed y descritte da joint *pdf*:

- l'evento A corrisponde a $x \in [x_0, x_0 + \mathrm{d}x]$ con y qualsiasi
- l'evento B corrisponde a $x \in [y_0, y_0 + \mathrm{d}y]$ con x qualsiasi

Le probabilità associate sono

$$P(A) = P(x \in [x_0, x_0 + \mathrm{d}x]) = pdf_{M_x}(x = x_0)\,\mathrm{d}x P(B) = P(y \in [y_0, y_0 + \mathrm{d}y]) = pdf_{M_y}(y = y_0)\,\mathrm{d}y$$

[rivedi]

Indipendenza. In questo caso si parla di indipendenza tra le due random variables che vale se e solo se si può scrivere

$$pdf(x, y) = pdf_{M_x}(x) \cdot pdf_{M_y}(y).$$

Il valore di aspettazione per una funzione $u(x, y)$ è

$$E[u(x, y)] = \iint u(x, y)\, pdf(x, y)\,\mathrm{d}x\,\mathrm{d}y.$$

Per media e varianza delle due variables

$$\mu_x = E[x] = \iint x\, pdf(x, y)\,\mathrm{d}x\,\mathrm{d}y$$
$$\mathrm{Var}[x] = E[(x - \mu_x)^2] = \sigma_x^2$$

Similmente per la y.

Covarianza. Si definisce la covarianza

$$\mathrm{Cov}[x, y] = E[(x - \mu_x)(y - \mu_y)] = E[x \cdot y)].\mu_x\mu_y.$$

Alcune proprietà

- se x e y sono indipendenti allora $\mathrm{Cov}[x, y] = 0$;
- se x e y sono legati linearmente allora $\mathrm{Cov}[x, y] = 0$;
- se $\mathrm{Cov}[x, y] = 0$ non implica che x e y sono indipendenti.

La matrice di covarianza si traduce in

$$\begin{pmatrix} \sigma_x^2 & \mathrm{Cov}[x, y] \\ \mathrm{Cov}[x, y] & \sigma_y^2 \end{pmatrix}.$$

Correlazione. Il coefficiente di correlazione è

$$\rho_{xy} = \frac{\mathrm{Covar}[x, y]}{\sigma_x \sigma_y}.$$

Alcune sue proprietà sono

- $\rho^2 \leq 1$
- se x e y sono legati linearmente allora $\rho^2 = 0$
- è sempre possibile operare un cambio di variabili che renda nulla la covarianza: si possono avere variabili non correlate ma che non sono indipendenti.

Due variabili non sono correlate quando la loro covarianza è nulla. Esse sono indipendenti quando $pdf(x, y) = pdf(x) \cdot pdf(y)$ che implica $\mathrm{Cov}[x, y] = 0$. La covarianza è una misura della dipendenza *lineare* delle variabili.
Le stime di parametri sono a loro volta random variables correlate tra loro, descritte da *pdf* congiunte.

Lecture 5

mar 02 nov 2021 13:30

4 Cambio di variabile in n-dimensioni, *pdf* normale multivariata, propagazione degli errori

4.1 Cambio di variabile

Si studia come cambia una *pdf* quando si opera una trasformazione di variabili

$$pdf_x(\vec{x}) : \mathbb{R}^n \to \mathbb{R}^+.$$

Si effettua un cambio di variabile $\vec{x} \to \vec{y} = \vec{f}(\vec{x})$. Si calcola l'espressione analitica della nuova *pdf*: $pdf_x(\vec{x}) \to pdf_y(\vec{y})$.

Caso monodimensionale Si è già vista la soluzione nel caso di una random variable monodimensionale: $pdf_x(x) : \mathbb{R} \to \mathbb{R}^+$. La funzione che descrive il cambio di variabile deve essere monotona, biunivoca e derivabile

$$x \to y = f(x) \quad \text{con inversa } g(y) = x.$$

L'espressione analitica della nuova *pdf* è

$$pdf_y(y) = pdf_x(x) \cdot |g'(y)| = pdf_x(x) \frac{1}{|f'(x)|}.$$

Caso bidimensionale. Siano x_1 e x_2 due random variables descritte con una *pdf* congiunta $pdf_x(x_1, x_2)$. Si applica una trasformazione di variabile (con funzioni monotone, biunivoche e derivabili):

$$\begin{cases} x_1 \to y_1 = u_1(x_1, x_2) \\ x_2 \to y_2 = v_2(x_1, x_2) \end{cases}.$$

Si cerca di capire qual è la relazione tra la joint *pdf* di partenza, $pdf_x(x_1, x_2)$, e quella di arrivo $pdf_y(y_1, y_2)$. Per analogia con quanto fatto nel caso monodimensionale, la $pdf_y(y_1, y_2)$ deve soddisfare la relazione:

$$pdf_y(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 = pdf_x(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2.$$

Scrivendo le funzioni inverse

$$\begin{cases} x_1 = w_1(y_1, y_2) \\ x_2 = w_2(y_1, y_2) \end{cases}.$$

Si ricava

$$pdf_y(y_1, y_2) = pdf_x(x_1, x_2) \cdot |J|.$$

La matrice J è la Jacobiana della trasformazione

$$|J| = \det \begin{bmatrix} \partial_{y_1} w_1 & \partial_{y_2} w_1 \\ \partial_{y_1} w_2 & \partial_{y_2} w_2 \end{bmatrix}.$$

La relazione si può scrivere come

$$pdf_y(y_1, y_2) = pdf_x(w_1(y_1, y_2), w_2(y_1, y_2)) \cdot |J|.$$

La trasformazione di variabili si usa per:

- ridefinire le variabili in modo che abbiano $\mathrm{Cov}[x_i, x_j] = 0$;
- la generazione di numeri casuali, come per esempio la trasformazione di Box-Muller: è un modo efficiente per generare numeri casuali con *pdf* normale partendo da `rand()`;
- ricavare la relazione di propagazione degli errori (visto successivamente).

4.2 *pdf* normale multivariata

Si è visto che quando un evento aleatorio è identificato da un vettore $\vec{x}$, si ha a che are con *pdf* multidimensionali, si parla anche di distribuzioni di probabilità multivariate. Con riferimento ad una variabile continua si può scrivere:

$$pdf(\vec{x}) : \mathbb{R}^n \to \mathbb{R}^+.$$

oppure, con notazione del tutto equivalente:

$$pdf(x_1, x_2, \ldots, x_n) : \mathbb{R}^n \to \mathbb{R}^+.$$

nell'ultimo caso è più frequente parlare di joint *pdf* delle n variabili. Per le *pdf* multivariate:

- il concetto di media della random variable monodimensionale è esteso ad una random variabile n-dimensionale in modo ovvio

$$\mathbb{E}[\vec{x}] = \vec{\mu} = \int \vec{x} \cdot pdf(\vec{x})\,\mathrm{d}\vec{x}.$$

- il concetto di varianza si estende introducendo una matrice di covarianza Σ definita da

$$\Sigma = \mathrm{Cov}[x_i, x_j] = \sigma_{ij}^2 = \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)].$$

Una distribuzione notevole è la distribuzione normale multivariata. Questa è un'estensione ad n-dimensioni della distribuzione gaussiana (o normale). Esattamente come la gaussiana monodimensionale, anche la gaussiana multivariata può essere parametrizzata in termini della propria media $\vec{\mu}$ e della propria matrice di covarianza Σ:

$$N(\vec{x};\vec{\mu},\Sigma) = \frac{1}{\sqrt{(2\pi)^n \cdot \det(\Sigma)}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^\top \cdot \Sigma^{-1} \cdot (\vec{x}-\vec{\mu})}.$$

In due dimensioni $\vec{x} = (x, y)$ si esplicitano le varianze e per la covarianza si utilizzano i coefficienti di correlazione:

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}.$$

Dunque segue

$$N(\vec{x};\vec{\mu},\Sigma) = \frac{1}{\sqrt{(2\pi)^2\sigma_x^2\sigma_y^2(1-\rho)}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}.$$

il termine dell'esponente può essere studiato per definire la forma delle curve di equiprobabilità. Dimostrazione negli appunti.
Per capire come sia fatta questa curva bidimensinale si possono determinare le curve di equiprobabilità che corrisponde alla richiesta che il termine all'esponente sia costante, K. Per semplificare l'espressione si può partire dalla gaussiana standard. Questa è detta distribuzione binormale standardizzata. Il massimo della funzione si ha quando $K = 0$, cioè all'origine. Quindi media e moda coincidono. Quindi

$$K = -\frac{1}{2(1-\rho^2)}\left[x^2 - 2\rho xy + y^2\right].$$

Si ha un caso particolare per cui l'esponente vale $K = -\frac{1}{2}$ ed è il valore per cui si riduce la densità di probabilità di un fattore $\frac{1}{\sqrt{e}}$ rispetto al massimo (che corrisponde a $K = 0$ e $\vec{\mu} = (0, 0)$. Nel caso monodimensinale la riduzione di un fattore $\frac{1}{\sqrt{e}}$ è quella che corrisponde all'intervallo $\mu \pm \sigma$ e che coincide con circa il 68% della probabilità. Nel caso bidimensionale questo intervallo è sostituito dall'ellissi degli errori. L'integrale della *pdf* su tale ellissi corrisponde ad una probabilità del 39%. L'ellissi degli errori è inscritto nel rettangolo definito dagli intervalli $\mu_x \pm \sigma_x$ e $\mu_y \pm \sigma_y$. L'integrale della binormale sia sull'ellissi (0.39) che sul rettangolo (0.47) restituisce una probabilità inferiore a quella a cui si fa riferimento per la gaussiana monodimensionale.

Diagonalizzazione di Σ. Un'importante proprietà della matrice di covarianza è il fatto che sia simmetrica. Infatti, le matrici simmetriche sono sempre diagonalizzabili mediante rotazione. La matrice che diagonalizza Σ definisce una trasformazione di variabili che consente di passare da variabili correlate a variabili non più correlate. Una rotazione di un angolo ϕ trasforma le coordinate secondo

$$\begin{aligned} x' &= x\cos\phi + y\sin\phi \\ y' &= -x\sin\phi + y\cos\phi \end{aligned}$$

E trasforma la matrice di covarianza secondo

$$\Sigma' = U\Sigma U^\top.$$

Se U è scelta in modo da essere la matrice che diagonalizza Σ, allora la rotazione fa in modo che le due variabili trasformate, x' e y', abbiano $\mathrm{Cov}[x', y'] = 0$. Dunque la *pdf* risultante è

$$pdf(x', y') = pdf(x, y) \cdot \left|\frac{\partial(w_1, w_2)}{\partial(x', y')}\right|.$$

dove le funzioni w sono quelle che descrivono il cambiamento di variabile

$$\begin{cases} x = w_1(x', y') \\ y = w_2(x', y') \end{cases}.$$

L'ultimo fattore è il determinante della matrice jacobiana

$$|J| = \det \begin{bmatrix} \partial_{x'} w_1 & \partial_{y'} w_1 \\ \partial_{x'} w_2 & \partial_{y'} w_2 \end{bmatrix}.$$

Nel caso considerato, il determinante della jacobiana è costante, dunque la *pdf* mantiene la stessa forma. Nel caso della binormale, la matrice di rotazione U che diagonalizza Σ è descritta da un angolo di rotazione ϕ:

$$\tan 2\phi = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}.$$

Il nuovo coefficiente di correlazione è $\rho' = 0$. Dunque

$$N(\vec{x}'; \vec{\mu}', \Sigma') = \frac{e^{-\frac{1}{2}\left[\left(\frac{x'-\mu'_x}{\sigma'_x}\right)^2 + \left(\frac{y'-\mu'_y}{\sigma'_y}\right)^2\right]}}{2\pi\sigma'_x\sigma'_y}.$$

Nelle variabili trasformate, la covarianza è nulla e $pdf(x', y') = pdf(x') \cdot pdf(y')$ cioè che le variabili sono indipendenti. Questa proprietà è solo della binormale.

5 Propagazione degli errori (statistici)

Si consideri il problema più generale possibile: determinare una grandezza fisica Y con misure indirette, sapendo che

$$Y = f(X_1, \ldots, X_n).$$

La procedura di misura è

- si misura $X_1, \ldots, X_n$ ottenendo i valori $x_1, \ldots, x_n$
- le quantità misurate sono descritte da una joint *pdf*: $pdf(x_1, \ldots, x_n)$
- per ogni set di dati $x_1, \ldots, x_n$ si ricava $y = f(x_1, \ldots, x_n)$
- la quantità $y = f(x_1, \ldots, x_n)$ è una random variable con $pdf(y)$ ignota.

Il problema da risolvere è di determinare $\mathbb{E}[y]$ e $\mathrm{Var}[y]$ che sono le stime per Y ed il suo errore. Come visto nel caso monodimensionale, $Y = f(X)$, il problema si può risolvere con due approcci:

- si determina per via analitica la $pdf(y)$
- si usa uno sviluppo di Taylor per calcolare $\mathbb{E}[y]$ e $\mathrm{Var}(y)$ in modo approssimato.

5.1 Caso analitico

Il problema da risolvere è quello di trovare $pdf(y)$ conoscendo la relazione tra le grandezze misurate direttamente e la grandezza indiretta a cui si è interessati. Il caso può apparire differente da quello risolto nel caso del cambiamento di variabile: si parte da $pdf(x_1, \ldots, x_n)$ e si deve arrivare a $pdf(y)$, cioè passare da n random variables ad una singola random variable. Si opera un cambiamento di variabili:

$$\begin{cases} x_1 \to y = f(x_1, \ldots, x_n) \\ x_2 \to x_2 \\ \vdots \\ x_n \to x_n \end{cases}.$$

A questo punto usando la regola della jacobiana si ricava la *pdf* espressa sulle nuove variabili $pdf'(y, x_2, \ldots, x_n)$ che poi si marginalizza (cioè si integra) su tutte le variabili che non interessano (cioè $x_2, \ldots, x_n$) ricavando $pdf(y)$:

$$pdf(y) = \int pdf(y, x_2, \ldots, x_n)\, \mathrm{d}x_2 \ \ldots\ \mathrm{d}x_n.$$

Da cui si ricava il valore di aspettazione e la sua varianza.

5.2 Caso approssimato

Si estende quando fatto per una dimensione. Se $y = ax+b$ allora si usano le proprietà di linearità dell'operatore del valore di aspettazione dimostrando quanto segue

$$\mu_y = \mathbb{E}[y] = a\mathbb{E}[x] + b = a\mu_x + b$$
$$\sigma_y^2 = \mathbb{E}[y^2] - \mu_y^2 = a^2\sigma_x^2$$

Se $y = y(x)$ è una funzione non lineare in x si può usare lo sviluppo di Taylor della funzione intorno a μ_x. Dunque risulta

$$\mu_y \approx y(\mu_x) + \frac{1}{2}\frac{\mathrm{d}^2 y(\mu_x)}{\mathrm{d}x^2}\sigma_x^2$$
$$\sigma_y^2 \approx \left(\frac{\mathrm{d}y(\mu_x)}{\mathrm{d}x}\right)^2 \sigma_x^2$$

L'ipotesi è di avere

- n variabili aleatorie $[x_1, \ldots, x_n]$ che possono essere rappresentate con un vettore $\vec{x}$
- una joint *pdf* che descrive le variabili che ha media $\vec{\mu}$ e matrice di covarianza con elementi σ_{ij}^2
- una variabile y che è una funzione delle n variabili: $y(\vec{x})$

La media risulta essere [rivedi] copiare

$$\mu_y = \int y(\vec{x})\, pdf(\vec{x})\, \mathrm{d}\vec{x} = \ldots \approx y(\vec{\mu}).$$

In modo analogo la varianza sviluppata al primo ordine è

$$\sigma_y^2 \approx \sum_{i,j=1}^{n} \partial_{x_i} y(\mu_x)\, \partial_{x_j} y(\mu_x) \sigma_{ij}^2.$$

Se le variabili x_i hanno correlazione nulla (quindi sicuramente quando esse sono indipendenti l'una dall'altra) allora la matrice di covarianza è diagonale e la regola di propagazione degli errori si riduce a

$$\sigma_y^2 \approx \sum_{i=1}^{n} (\partial_{x_i} y(\mu_x))^2 \sigma_{ij}^2.$$

Risulta essere utile cosa succede nei casi semplici. Da queste formule discendono le due semplici regole per cui nella somma di variabili gli errori si propagano gli errori sommandoli in quadratura; mentre nei prodotti si sommano in quadratura gli errori relativi.

Lecture 6

gio 11 nov 2021 13:30

6 Random walk in due dimensioni

Si consideri una particella che si muove in un piano incontrando ostacoli che ne modificano il moto in modo non deterministico. Si consideri un caso semplice, che ha soluzione analitica:

- la particella compie passi di lunghezza fissa $L = 1$
- dopo ciascun passo, la particella si muove in una nuova direzione descritta da un angolo θ con $pdf(\theta)$ uniforme.

Si vuole sapere dove si trova la particella dopo N passi e in particolare con che probabilità essa si trovi in un punto (x, y) del piano.

Parte 1. Si vuole capire cosa succede al primo passo. Si suppone che la particella parta dall'origine del sistema di coordinate. Ci si chiede qual è la distribuzione di probabilità della particella dopo il primo passo.
La particella vive nel piano xy. Parte dall'origine e compie un passo di lunghezza unitaria con un angolo θ. La distribuzione degli angoli è uniforme: $pdf(\theta) = \frac{1}{2\pi}$. Si possono anche utilizzare le coordinate cartesiane:

$$\begin{cases} x = \cos\theta \\ y = \sin\theta \end{cases}.$$

Esse sono due random variables perché funzioni di una random variable. Dunque, esiste

$$pdf(x, y) = \begin{cases} 0, & x^2 + y^2 \neq 1 \\ \text{cost.}, & x^2 + y^2 = 1 \end{cases}.$$

Tale costante si calcola in modo che l'integrale sul sample space sia 1. Dunque, la costante vale $\frac{1}{2\pi}$. Tuttavia, si vorrebbe calcolare la $pdf(x)$; per ottenerla si può immaginare di fare una trasformazione di variabili:

$$\begin{cases} x = \cos\theta \\ pdf(\theta) = \frac{1}{2\pi} \end{cases}.$$

Si può calcolare la pdf marginale, tuttavia è più semplice utilizzare un cambio di coordinate. Dunque

$$pdf(x)\,\mathrm{d}x = pdf(\theta)\,\mathrm{d}\theta \implies pdf(x) = |\partial_x\theta|\, pdf(\theta).$$

Le funzioni per il cambio di variable non sono strettamente monotone sull'intervallo $[-\pi, \pi]$, questo perché la regola della trasformazione richiede che siano monotone. Tuttavia, basta fare il conto per tratti.
A questo punto

$$\begin{cases} \theta = \operatorname{acos} x \\ \theta = \operatorname{asin} y \end{cases}.$$

Pertanto

$$pdf_M(x) = \frac{1}{2\pi}\frac{1}{\sqrt{1 - x^2}}.$$

La distribuzione per y è identica. Si vogliono trovare i momenti. Non si fanno gli integrali su x, ma si ritorna a θ. Infatti

$$\mathbb{E}[u(\theta)] = \int pdf(\theta)\, u(\theta)\,\mathrm{d}\theta.$$

Dunque, posto $x = \cos\theta$ si ha

$$\mathbb{E}[x] = \int_{-\pi}^{\pi} \frac{1}{2\pi}\cos\theta\,\mathrm{d}\theta = 0.$$

Allo stesso modo

$$\mathbb{E}[x^2] = \int_{-\pi}^{\pi} \frac{1}{2\pi} \cos^2\theta \, \mathrm{d}\theta = \frac{1}{2}.$$

Da cui la varianza risulta essere

$$\mathrm{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \frac{1}{2}.$$

Per la variabile y è identico. Per la *pdf* congiunta

$$E[x] = \int x \, pdf(x,y) \, \mathrm{d}x \, \mathrm{d}y = \int x \, \mathrm{d}x \underbrace{\int pdf(x,y) \, \mathrm{d}y}_{pdf_M(x)} = \int x pdf_M(x) \, \mathrm{d}x.$$

che risulta essere quanto visto fin'ora. [rivedi] Quindi $E[x] = E[y] = 0$. La matrice di covarianza risulta essere

$$\Sigma = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} = \begin{bmatrix} \mathrm{Var}[x] & \mathrm{Cov}[xy] \\ \mathrm{Cov}[yx] & \mathrm{Var}[y] \end{bmatrix}.$$

La covarianza è $E[(x-\mu_x)(y-\mu_y)] = \mathrm{Cov}[xy] = E[xy] = 0$ questo perché la distribuzione è un cerchio che è una figura per cui la covarianza è nulla. Infatti, tali due variabili non sono linearmente correlate tra loro, tuttavia non sono indipendenti perché entrambe dipendono da θ. Quindi

$$\mathrm{Cov}[xy] = E[xy] = \int xy \, pdf(x,y) \, \mathrm{d}x \, \mathrm{d}y = \int r \cos\theta \sin\theta \, \mathrm{d}r \, \mathrm{d}\theta?.$$

[rivedi] Dopo N passi si ha

$$\begin{cases} x_N = \sum_{i=1}^{N} \cos\theta_i \\ y_N = \sum_{i=1}^{N} \sin\theta_i \end{cases}.$$

Dunque

$$E[x_N] = E[\sum x_i] = \sum E[x_i] = 0.$$

Mentre

$$E[x_N^2] = \frac{N}{2}, \quad \mathrm{Var}[x] = \frac{N}{2}.$$

Pertanto

$$pdf(x_N) = .$$

[rivedi] Tuttavia, il comportamento asintotico è una gaussiana per il teorema centrale del limite perché x_N è una variabile casuale somma di variabili casuali:

$$G(x;\mu,\sigma), \quad \mu = 0, \quad \sigma = \frac{N}{2}.$$

[rivedi] Si dimostra che $\mathrm{Cov}[x_N y_N] = 0$, dunque anche dopo N passi le variabili non sono correlate linearmente. Inoltre, la distribuzione per $N \to \infty$ è una binormale. Essa ha una particolarità: cioè la covarianza è nulla. Questo significa che le due variabili sono indipendenti e

$$pdf(x_N, y_N) = pdf(x_N) \, pdf(y_N) = \frac{1}{2\pi\sigma_N} e^{-\frac{x^2}{2\sigma_N}} e^{-\frac{y^2}{2\sigma_N}}.$$

Può essere utile passare alle coordinate polari per cui si ottiene

$$pdf(x_N, y_N) = |J| \, pdf(\rho, \varphi).$$

Ci si aspetta che la *pdf* dipenda solamente da ρ perché il problema è simmetrico in φ e si arriva a

$$pdf(\rho) = k\rho e^{-\frac{\rho}{2\sigma_N}}.$$

Che non è una gaussiana a causa del fattore ρ.

Lecture 7

gio 18 nov 2021 13:30

7 Statistica

La probabilità è un aleatoria matematica che regola il comportamento di una variable aleatoria (r.v.). Le basi sono

- gli assiomi di Kolmogorov e la definizione di *pdf*
- il concetto di popolazione, campione e campionamento
- i parametri descri [rivedi]

In fisica le random variables sono usate per la descrizione dei fenomeni:

- intrinsecamente non deterministici, perché regolati dalla meccanica quantistica
- troppo complessi per poter essere descritti in modo deterministico

La statistica è la teoria che studia il problema inverso della probabilità: dato un campione (i dati sperimentali) si vuole inferire le proprietà ed i parametri (inclusa la *pdf*) che descrivono la popolazioni dalla quale il campione è estratto.

- Si inferisce il valore di un parametro (detto stimatore).
- Verificare o confutare la validi di un modello (cioè il test di ipotesi).

7.1 Independent Identically Distributed

Dati indipendenti ed identicamente distribuiti. Si dice che N campionamenti sono IID quando sono

- indipendenti — l'esito di un campionamento non è influenzato da nessuno degli altri
- identicamente distribuiti — quindi estratti dalla stessa $pdf_x(x) : \mathbb{R} \to \mathbb{R}^+$.

La joint *pdf* di N campionamenti IID è il prodotto delle singole probabilità in ragione dell'indipendenza:

$$pdf_{\text{set}}(x_1, \dots, x_N) = \prod_{i=1}^{N} pdf_x(x_i, \theta).$$

Il parametro θ è l'insieme dei parametri che descrivono la *pdf*. Pertanto la *pdf* set misura la probabilità di estrarre un particolare set di dati ed è una funzione definita su uno spazio N-dimensionale.

Statistica. Si definisce statistica una funzione di N campionamenti IID che contenga solo parametri noti.
Una statistica è una variabile aleatoria. QUindi ha una sua pdf_f derivabile dalla joint *pdf* dei campionamenti e dalla forma della funzione f. Si hanno 3 *pdf*:

- la pdf_x che si campiona IID
- pdf_{set} dei campionamenti (che è multidimensionale)
- la pdf_f della statistica dei campionamenti

Parametri da stimare. Quando non si conoscono i parametri che descrivono una pdf_x, la procedura che si può adottare è di usare N campionamenti IID per estrarre le informazioni desiderate. Essendo basate su di un campione, queste informazioni non sono esatte, ma sono delle stime, provviste di una loro incertezza. Si può voler stimare

- la media $\mu = E[x]$ e la varianza $\sigma^2 = \mathrm{Var}[x]$ della pdf_x
- in generale, i parametri che descrivono la distribuzione $pdf_x(x, f)$

Stimatori. Uno stimatore è una statistica opportunamente scelta in modo da poter usare N campionamenti IID per stimare il valore di uno più parametri descrittivi della pdf_x.

Notazione. Si vedono alcune notazioni

- la pdf il cui parametro θ va stimato (spesso chiamato θ_V valore vero del parametro) è $pdf_x(x, \theta)$
- i campionamenti IID della pdf_x sono $\{x_1, \ldots, x_N\}$
- la statistica che definisce lo stimatore è $\hat{\theta}(x_1, \ldots, x_N)$
- la pdf della random variable stimatore è $pdf_{\hat{\theta}}(\hat{\theta})$
- il valore dello stimatore ottenuto per uno specifico campionamento è $\hat{\theta}^*$

Consistenza. Lo stimatore è detto consistente se per $N \to \infty$ lo stimatore restituisce il valore vero del parametro. Quando il campione si avvicina all'intera popolazione lo stimatore deve essere esatto.

Valore di aspettazione di uno stimatore. Uno stimatore, in quanto random variabile, è caratterizzato da una $pdf_{\hat{\theta}}(\hat{\theta})$ e quindi ha un valore medio ed una varianza.

$$E[\hat{\theta}] = \int \hat{\theta}\, pdf_{\hat{\theta}}(\hat{\theta})\, \mathrm{d}\hat{\theta} = \int \ldots \in \hat{t}\theta(x_1, \ldots, x_N)\, pdf_x(x_1) \ldots pdf_x(x_N)\, \mathrm{d}x_1 \ldots \mathrm{d}x_N.$$

Le relazioni precedenti sfruttano due proprietà viste:

- la conservazione della probabilità in un cambiamento di variabile $y = y(x) : pdf(x)\, \mathrm{d}x = pdf(y)\, \mathrm{d}y$
- la regola per la riduzione del numero di variabili indipendenti $x_1, \ldots, x_N \to \theta$ in cui la $pdf(\theta)$ si ottiene cambiando le variabili e marginalizzando

Le qualità dello stimatore possono essere definite per confronto tra questi due parametri ed il valore vero. Lo stimatore è tanto più accurato (o non-distorto, unbiased) quanto più vicino è il uso valore medio $E[\hat{\theta}]$ al valore vero. Lo stimatore è tanto più efficiente quanto minore è la sua varianza $\mathrm{Var}\left[\hat{\theta}\right]$ (partendo da set di campionamenti differenti si desiderano avere stime simili). Le proprietà di uno stimatore sono

- la consistenza; se $N \to \infty$ allora $\hat{\theta}_N \to \theta_{\mathrm{vero}}$
- il bias, $b_N = E[\hat{\theta}_N] - \theta_{\mathrm{vero}}$; se $b_N = 0$ si dice che lo stimatore è unbiased o non distorto. Questo si traduce nell'accuratezza della misura.
- l'efficienza, misurata da $\mathrm{Var}\left[\hat{\theta}_N\right]$ ha un valore minimo fissato N che è definito dal teorema di Rao-Kramer. Questo si traduce nella precisione della misura.

Esistono tecniche di costruzione di stimatori che permette di sapere a priori che lo stimatore è privo di bias.
Nella realtà, molte volte la misura vera non è nota, dunque non risulta immediato capire se esiste un bias o meno.
Per uno stesso parametro si possono definire tanti stimatori, ma non tutti hanno le proprietà desiderate. Va anche notato che non è detto che esista (o si sappia trovare) uno stimatore che soddisfi contemporaneamente tutte le richieste.

Stima della media e della varianza. La procedura di stima coinvolge

- trovare uno stimatore $\hat{\theta}$
- valutare la consistenza, il bias e la varianza
- se possibile calcolare la sua $pdf_{\hat{\theta}}(\hat{\theta})$

Media campionaria. Lo stimatore per μ che meglio soddisfa i requisiti ideali è la media campionaria (sample mean) spesso indicata con $\overline{x}$

$$\hat{\mu}_N = \overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_i.$$

Si vede che

- è unbiased, $E[\overline{x}] = \mu$
- è efficiente, con una varianza che decresce all'aumentare del numero di campionamenti: $\mathrm{Var}\,[\overline{x}] = \frac{\sigma^2}{N}$
- è consistente se la $pdf_x(x)$ soddisfa le ipotesi del teorema centrale del limite
- la $pdf_{\overline{x}}(\overline{x})$ è gaussiana per $N \to \infty$ se la $pdf_x(x)$ soddisfale ipotesi del teorema centrale del limite.

Si dimostra che

$$E[\overline{x}] = \int \overline{x}\, pdf_{\overline{x}}(\overline{x})\, \mathrm{d}\overline{x} = \mu$$
$$\mathrm{Var}\,[\overline{x}] = \frac{1}{N}\sum_{n=1}^{N} \mathrm{Var}\,[x_i] = \frac{\sigma^2}{N}$$

Quindi lo stimatore è unbiased e all'aumentare del numero di campionamenti la sua varianza si riduce. Questo garantisce che per un alto numero di campionamenti, le stime ottenute con diversi set di dati sono vicine tra loro. Inoltre, essendo vicine alla media dello stimatore, che è unbiased, ne discende che le stime sono vicine al valle vero μ.

Distribuzione e consistenza. Valgono le seguenti proprietà

- se $pdf_x(x)$ è gaussiana allora anche $pdf_{\overline{x}}(\overline{x})$ lo è in quanto $\overline{x}$ è somma di random variables gaussiane e per la gaussiana vale la proprietà di riproducibilità.
- se la $pdf_x(x)$ non è gaussiana, ma soddisfa le ipotesi del teorema centrale del limite, allora $pdf_{\overline{x}}(\overline{x}) \to$ Gaussiana quando $N \to \infty$.

In entrambi i casi

$$\mathrm{Gauss}\left(\overline{x}; \mu, \frac{\sigma}{\sqrt{N}}\right).$$

e lo stimatore è consistente.

Stimatore per la varianza. Lo stimatore per la varianza che meglio soddisfa i requisiti ideali è la varianza campionaria (sample variance). Sono possibili due definizioni

- lo scarto quadratico medio rispetto alla media della popolazione μ nota a priori

$$s_\mu^2 = \frac{1}{N}\sum_{n=1}^{N}(x_i - \mu)^2.$$

  molte volte μ non è noto e pertanto bisogna usare la definizione seguente.

- lo scarto quadratico medio rispetto alla media del campione $\overline{x}$

$$s_{\overline{x}}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_i - \overline{x})^2.$$

Si dimostra che

- il bias di s_μ^2 è nullo: $b_N = E[s_\mu^2] - \sigma^2 = 0$
- il bias di $s_{\overline{x}}^2$ è diverso da zero

$$b_N = E[s_{\overline{x}}^2] - \sigma^2 = \left(\frac{N-1}{N}\sigma^2\right) - \sigma^2.$$

Risulta chiaro che il primo stimatore, pure essendo unbiased, è poco utile perché è raro che si conosca μ, ma non σ^2. Il secondo stimatore è quello normalmente adottato, è però distorto anche se asintoticamente è unbiased.
Si può definire un terzo stimatore che introduca una correzione a $s_{\overline{x}}^2$ tale da cancella il bias. La correzione del bias è applicabile tutte le volte in cui il bias sia precisamente noto. Il nuovo stimatore della varianza è

$$s^2 \equiv \frac{1}{N-1}\sum_{n=1}^{N}(x_i - \overline{x})^2.$$

Tale stimatore è privo di bias: $E[s^2] = \sigma^2$ e la varianza non può essere determinata per il caso generale

$$\mathrm{Var}\left[s^2\right] = \mathrm{Var}\left[\frac{1}{N-1}\sum_{n=1}^{N}(x_i - \overline{x})^2\right] = \frac{1}{(N-1)^2}\,\mathrm{Var}\left[\sum_{n=1}^{N}(x_i - \overline{x})^2\right].$$

mentre è possibile farlo per il caso particolare in cui $pdf_x(x)$ sia una gaussiana. Se si riscrive la random variable che definisce la varianza campionaria in questo modo

$$s^2 \equiv \frac{1}{N-1}\sum_{n=1}^{N}(x_i - \overline{x})^2 = \frac{\sigma^2}{N-1}\sum_{n=1}^{N}\frac{(x_i - \overline{x})^2}{\sigma^2}.$$

si può introdurre una random variable ausiliare chiamato χ^2

$$s^2 \equiv \frac{\sigma^2}{N-1}\chi^2.$$

Inoltre, $\mathrm{Var}\left[s^2\right]$ è legata a $\mathrm{Var}\left[\chi^2\right]$.
Nell'ipotesi in cui si stia campionando una pdf_x gaussiana, segue che la variabile χ^2 ha una *pdf* nota che si chiama distribuzione chi-quadro e la sua random variable è χ^2. La *pdf* chi-quadro è descritta da un solo parametro chiamato gradi di libertà. Nel caso ora analizzato il parametro vale $N-1$. Quindi

$$s^2 = \frac{\sigma^2}{N-1}\chi^2_{N-1}.$$

Si possono sfruttare le proprietà note della distribuzione chi-quadro:

$$E[\chi^2_N] = N \implies E[s^2] = \frac{\sigma^2}{N-1} E[\chi^2_{N-1}] = \frac{\sigma^2}{N-1}(N-1) = \sigma^2$$
$$\text{Var}\left[\chi^2_N\right] = 2N \implies \text{Var}\left[s^2\right] = \frac{\sigma^4}{(N-1)^2}\,\text{Var}\left[\chi^2_{N-1}\right] = \frac{\sigma^4}{(N-1)^2} 2(N-1) = \frac{2\sigma^4}{(N-1)}$$

Lecture 8

gio 25 nov 2021 13:30

La domanda tipica dell'esame è discutere il problema della stima, spiegando cos'è uno stimatore, le sue proprietà [rivedi] Bisogna introdurre il contesto in cui si pone il problema: si vuole misurare una grandezza, ma i dati sono campionamenti $\{x_1, \ldots, x_n\}$ di una *pdf* e la grandezza che si vuole misurare è uno dei parametri descrittivi della *pdf*.
Questo può succedere quando:

- la misura è affetta da errori casuali. Quando si misura, si campiona una gaussiana con media sul valore vero e sigma che dipende dagli strumenti di misura. Quindi si ha una distribuzione *pdf* $\left(x, \vec{\theta} = \begin{pmatrix}\mu \\ \sigma\end{pmatrix}\right)$.
- si misura un fenomeno intrinsecamente aleatorio; cioè tutte le volte in cui la descrizione di quanto si vuole misurato è definito dalla meccanica quantistica. La grandezza non è definita da un valore, ma da una distribuzione di probabilità dei valori. Le righe di spettro di emissione ha una distribuzione *pdf*$(x, \vec{\theta})$ con $\vec{\theta}$ composto dai valori λ_i e σ_i per i indice che identifica una linea di spettro. I parametri della *pdf* non sono direttamente i singoli λ e σ anche se ad essi sono legati.

Si studia cos'è la stima tramite l'esempio primo. Si supponga di fare una misura x, allora si riporta come $\theta^* = x \pm \sigma$. La stima di μ è legata a quanto misurato x e ad una incertezza σ legata alla gaussiana, associata ad un confidence level: esso è la probabilità che il valore vero sia all'interno dell'intervallo $[x - \sigma, x + \sigma]$.
Ci si aspetta che la stima si possa scrivere come $\theta^* \pm \delta\theta^*$. Inoltre, si vuole sapere quanto questo intervallo sia vicino al vero θ_V; si vuole sapere quant'è grande l'errore; ed in generale sapere com'è fatta la distribuzione di probabilità che definisce θ^*.
La soluzione del problema è legata all'utilizzo dei campionamenti. Si costruisce una funzione dei campionamenti $f(x_1, \ldots, x_n)$ e tale si è chiamata statistica. Si vuole la statistica (anch'essa random variable) che determina il parametro θ della *pdf* e la si chiama $\hat{\theta}(x_1, \ldots, x_n)$. Di questa random variable [rivedi] La random variable ha una distribuzione *pdf*$(\hat{\theta}, \vec{\theta})$. [rivedi]
Quando si ha $\hat{\theta}^*$ si ha un singolo campionamento della *pdf*.
Si vuole capire quanto la stima sia vicina al valore vero. [rivedi]

Si vuole sapere se $E[\hat{\theta}] = \theta_V$ e questo è ciò che si è chiamato bias (accuratezza). Inoltre, si vuole sapere Var $\left[\hat{\theta}\right]$ (precisione). La stima è un singolo campionamento della *pdf*. Si vuole trovare una statistica che definisca la stima con una varianza piccola.

Due osservazioni. Si consideri di usare sempre la stessa procedura di campionamento e lo stesso stimatore. Ogni volta si ha una stima diversa, ma con la stessa incertezza. Si consideri di operare una misura con una tecnica diversa: essa potrebbe essere più precisa ma avere un bias. Questo succede quando si confrontano misure ottenute con metodi diversi. Tali valori potrebbero essere tra loro compatibili o meno. Bisogna dedurre che una delle due tecniche possiede un bias.
Si hanno tre casi:

- non si sta misurando la stessa cosa (il modello è sbagliato)
- in un esperimento si è sottostimato l'errore

- una delle due tecniche ha un bias

Precedentemente si è analizzato un caso semplice di costruzione di uno stimatore: stimare la media di una *pdf*. Lo stimatore della media è

$$\overline{x} = \frac{\sum x_i}{N}.$$

e si è dimostrato $E[\overline{x}] = \mu$, quindi non ha bias. Similmente Var $\left[\frac{\sigma^2}{N}\right]$
Si hanno campioni $\{x_1, \dots x_n\}$ con $pdf(x, \vec{\theta})$ e $E[x] = \mu$, Var $[x] = \sigma$. [rivedi]

Se la distribuzione dello stimatore soddisfa le ipotesi del teorema centrale del limite allora per $N \to \infty$ la distribuzione è una gaussiana centrata in μ di varianza $\frac{\sigma^2}{N}$. All'infinito la gaussiana collassa nella delta di Dirac centrata in μ. La condizione all'infinito è considerare tutta la popolazione: la proprietà per la quale all'infinito lo stimatore restituisce il valore vero è detta consistenza.
Quindi la stima è

$$\overline{x} \pm \frac{\sigma}{\sqrt{N}}.$$

dato che si ritorna ad una gaussiana, segue che l'intervallo è al 68%. Tuttavia, σ potrebbe non essere noto. Dunque, si costruisce uno stimatore anche per la varianza.

Stimatore varianza:

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{N - 1}.$$

essa è la varianza campionaria con correzione di Bessel. Si è visto che $E[s^2] = \sigma^2$ e non ha bias. Se la *pdf* è una gaussiana allora segue che

$$\text{Var}\left[s^2\right] = \frac{2\sigma^2}{N - 1}.$$

dato che, se si considera la distribuzione originale una gaussiana, allora la *pdf* di s^2 è una distribuzione chi-quadro con $N - 1$ gradi di libertà, per cui si sa calcolare la varianza.
Per lo stimatore della varianza, senza Bessel, si avrebbe ottenuta una distribuzione con bias. Applicando Bessel, si rimuove il bias, e nelle ipotesi della gaussiana si può calcolare la varianza. Così si sa sia la varianza che la forma della distribuzione.
Il problema non è del tutto risolto. Se si pone la stima come $\overline{x} \pm \frac{s}{\sqrt{N}}$ si hanno due variabili descritta da due *pdf*. [rivedi]
Il problema degli stimatori si riassume nel definire un intervallo e sapere quale probabilità tale intervallo contiene il valore vero: questa è la costruzione degli intervalli di confidenza. Mettendo s al posto di σ si ottiene la distribuzione di Student. [rivedi]

Si analizza il problema generico di capire come si trova uno stimatore $\hat{\theta}(x_1, \dots, x_n)$, trovare i suoi valor medio, varianza e *pdf* dove possibile. Si vede il metodo del maximum likelihood.

7.2 Distribuzione chi-quadro

La distribuzione χ^2 permette di costruire il test di ipotesi. Si costruisce una random variable χ^2 estraendo N valore IID, ciascuno da una pdf_i gaussiana di media μ_i e varianza σ_i^2 con

$$\chi^2 = \sum \frac{(x_i - \mu_i)^2}{\sigma_i^2}.$$

dove N è detto numero di gradi di libertà ed è 'unico parametro che determina la forma della *pdf* del χ^2.
La joint-*pdf* dei campionamenti è una *pdf* normale multivariata in uno spazio N-dimensionale (con matrice di covarianza Σ diagonale per l'ipotesi IID). Si può pensare al set di campionamenti

come un singolo campionamento (vettoriale) di una distribuzione normale N-dimensionale.
Se esistono n relazioni che legano tra loro le x_i allora si sta campionando una normale multivariata che ha matrice di covarianza non diagonale. Il numero di gradi gradi di libertà da associare alla random variable χ^2 è in questo caso $N - n$. Questo è il caso considerato per la variabile s^2 (la varianza campionaria) in cui si campiona una stessa pdf_x gaussiana. La variabile χ^2 è definita come

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - \overline{x})^2}{\sigma^2}.$$

ha $N - 1$ gradi di libertà perché esiste una relazione che lega gli x_i tra loro: $\frac{1}{N}\sum_{i=1}^{N} x_i = \overline{x}$.
La *pdf* random variable χ^2_N dipende da un solo parametro N ed è chiamata distribuzione del chi-quadro ad N gradi di libertà. La sua forma analitica è

$$pdf(\chi^2_N) = (\chi^2)^{\frac{N}{2}-1} \frac{e^{\frac{\chi^2}{2}}}{\Gamma\left(\frac{N}{2}\right)} 2^{\frac{N}{2}}.$$

dove Γ è la funzione gamma. La media è $E[k] = N$, la varianza $\mathrm{Var}[k] = 2N$, vale la riproduttività, la moda è $N - 2$ (e vale 0 per $N = 2$). Per $N > 2$ la *pdf* è mono-modale e si sposta verso destra più gradi di libertà si hanno; inoltre, diventa sempre più simmetrica.
Inoltre, si spesso definisce il χ^2 ridotto che è il rapporto $\frac{\chi^2_N}{N}$ la cui media è 1.
Nel caso particolare in cui le $pdf(x_i)$ sono gaussiane che derivano da distribuzioni binomiali o poissoniane, quindi descritte da un solo parametro μ_i (essendo $\sigma_i^2 = \mu_i$). La variabile χ^2_N assume la forma

$$\chi^2_N = \sum_{i}^{N} \frac{(x_i - \mu_i)^2}{\mu_i}.$$

Ponendo $\mu_i = E_i$ e $x_i = O_i$ allora si ha

$$\chi^2_N = \sum_{i}^{N} \frac{(O_i - E_i)^2}{E_i}.$$

Questo vale solo per eventi di conteggio.

7.3 Likelihood

La funzione di verosimiglianza $\mathcal{L}$ è definita come la probabilità di osservare il campione di dati IID, $\vec{x} = \{x_1, \ldots, x_N\}$, condizionata al valore assunto dal parametro θ oggetto di stima. Il funzionale $\mathcal{L}(\vec{x}, \theta)$ si può scrivere solo se è nota la forma analitica della $pdf_x(\vec{x}, \theta)$:

$$\mathcal{L}(\vec{x}, \theta) = \prod pdf_x(x_i, \theta).$$

la variabile indipendente è θ.

Utilizzi. La likelihood può essere utilizzata in diversi modi:

- per misurare l'informazione che i campionamenti contengono relativamente al parametro. Questo consente di studiare tecniche ottimali per la riduzione dei dati senza perdita di informazioni.
- per valutare la minima varianza raggiungibile da uno stimatore: dati i campionamenti, e quindi la loro likelihood, esiste un limite inferiore alla varianza di qualsiasi stimatore, oltre quel limite non è possibile andare.
- per definire un metodo con cui costruire uno stimatore.

Disuguaglianza di Rao-Cramer. Detta anche minimum variance bound (MVB). Si dimostra che esiste la seguente relazione tra la varianza di un qualsiasi stimatore $\hat{\theta}$ e la funzione di likelihood $\mathcal{L}(\vec{x};\theta)$:

$$\mathrm{Var}\left[\hat{\theta}\right] \geq \frac{[1+\partial_\theta b_n]}{E\left[-\partial_\theta^2 \ln \mathcal{L}(\vec{x};\theta)\right]}.$$

dove b_N è il bias dello stimatore considerato. Si vedere che il termine al denominatore fornisce una misura dell'informazione. Allora il MVB afferma che la varianza dello stimatore non può andare oltre un valore che è inversamente proporzionale all'informazione contenuta nel campione di dati. Uno stimatore è efficiente se vale si segno di uguaglianza, se tale stimatore esiste.

Maximum likelihood. la tecnica della massima verosimiglianza identifica lo stimatore per il parametro θ come il valore $\hat{\theta}_{\mathrm{ML}}$ in corrispondenza del quale la probabilità associata al campionamento $\mathcal{L}(\vec{x},\theta)$ è quella maggiore possibile.

$$\mathcal{L}(\theta) = \prod pdf_x(x_i,\theta).$$

lo stimatore $\hat{\theta}_{\mathrm{ML}}$ della maximum likelihood è quello per cui $\mathcal{L}(\theta)$ raggiunge il suo massimo assoluto.
Poiché $\mathcal{L}$ è un prodotto di N fattori segue che è pratico considerare il suo logaritmo che diventa un a sommatoria di N termini ed è detta log-likelihood:

$$\ln \mathcal{L}(\vec{x},\theta) = \sum \ln pdf_x(x_i,\theta).$$

In quanto il logaritmo è una funzione monotona, il massimo di $\mathcal{L}(\theta)$ è anche

- il massimo di $\ln \mathcal{L}(\theta)$
- il minimo di $-\ln \mathcal{L}(\theta)$

Nel caso di un parametro monodimensionale, lo stimatore ML è tale che

$$\partial_\theta \mathcal{L}(\hat{\theta}_{\mathrm{ML}}) = 0.$$

condizione a cui va aggiunto che l'estremante trovato sia massimo assoluto. In modo analogo, per il caso di un parametro pluridimensionale $\vec{\theta}$ lo stimatore della ML è il valore per cui si annullano tutte le derivate parziali.

Proprietà. Lo stimatore della ML può essere distorto (cioè bias diverso da zero) anche se è possibile dimostrare che, sotto ipotesi molto generiche, è asintoticamente unbiased quando il numero di campionamenti tende ad infinito. Similmente, si dimostra che gli stimatori ML sono asintoticamente efficienti. Inoltre, si verificare che se esiste uno stimatore efficiente, allora lo stimatore ML lo è a sua volta.

Lecture 9

gio 02 dic 2021 13:30

[r] tutto, cfr slide
Si son fatte delle misure $\{x_1,\ldots,x_N\}$. Si vuole stimare il parametro θ al meglio. Si costruisce uno stimatore senza bias, con varianza piccola e con consistenza (cioè che quando la stima è fatta su tutta la popolazione, allora lo stimatore restituisce il valore vero).
Si parte dall'unica cosa si può scrivere relativamente ai campionamenti IID. Si definisce la joint-*pdf* chiamata likelihood e dato che sono indipendenti, essa si può calcolare come

$$\mathcal{L}(x_1,\ldots,x_N;\theta) = \prod_{i=1}^{N} pdf(x;\theta).$$

Se θ è variabile, allora si ha una probabilità condizionata. Essa dice quant'è la probabilità di ottenere un set di campionamenti se θ ha un determinato valore. Ci si aspetta che se θ è il valor vero allora la probabilità è massima. Definendo la likelihood come funzione di θ, allora il massimo si presente nel valore vero.
Inoltre, si definisce l'informazione. Fissati i campionamenti, si ha un limite a quanto bello sia il parametro. La qualità dello stimatore è legata ai campionamenti tramite l'informazione.
Costruendo stimatori differenti con uno stesso set di dati, si arriva allo stimatore migliore quando la disuguaglianza di Rao-Cramer diventa un'identità, cioè la MVB è minima.
Costruendo la *pdf* dello stimatore, il bias misura il valore di aspettazione dello stimatore. Uno stimatore è una funzione delle misure; esso è una random variable, dunque si può disegnare una *pdf*.
La likelihood in funzione delle x è una joint-*pdf*. La sua derivata calcolata sui valori campionati è un numero. Essa è una funzione del parametro. Infatti, per una sola random variable x, descritta da una $pdf(x)$ si ha

$$E[\mu(x)] = \int_\Omega \mu(x)\, pdf(x)\, \mathrm{d}x.$$

tuttavia, ora si hanno N variabili casuali, con una joint-*pdf* corrispondente alla likelihood, dunque si calcolare il valore aspettazione di $\mu(\ldots) = \partial_\theta^2 \mathcal{L}$. La derivata seconda è la misura della concavità.
La varianza dello stimatore è un numero, quindi pure il numeratore e denominatore in Rao-Cramer. Dunque, il valore di aspettazione della derivata seconda della likelihood dev'essere il valore vero ed esso è detto informazione.?
Il modo in cui si formalizza il passaggio dalla likelihood a Rao-Cramer è tramite la definizione dell'informazione $I(\theta)$.
Dato un certo numero di campionamenti IID $\{x_1, \ldots, x_N\}$, si misura l'informazione contenuta in tali campionamenti. In generale, per un'informazione si vuole che il suo valore sia nullo se i dati sono irrilevanti per la stima del parametro; si vuole che cresca all'aumentare del numero di dati rilevanti; si vuole avere una relazione tra $I(\theta)$ e Var $\left[\hat{\theta}\right]$.
Si vuole separare il modo in cui si costruisce lo stimatore dalla varianza. Il modo in cui si costruisce l'informazione di Fischer parte dalla likelihood. Si parte da un singolo campionamento.

Score per un campionamento. Si consideri un unico campionamento per la $pdf(x;\theta)$. Si consideri che la *pdf* è gaussiana ed abbia σ noto e μ ignoto. Dato che non si conosce μ si può pensare alla distribuzione come una distribuzione condizionata. Si ha

$$\Delta p = \partial_\theta pdf(x,\theta)\Delta\theta.$$

Siccome la stessa variazione di probabilità ha un differente impatto a seconda che per quel valore di θ la probabilità sia lata o sia bassa. Si passa a considerare la variazione relativa della probabilità:

$$\frac{\Delta p}{p} = \frac{1}{pdf_x(x;\theta)} \partial_\theta pdf_x(x;\theta) \frac{\Delta\theta}{\theta}.$$

il fattore moltiplicativo di $\frac{\Delta\theta}{\theta}$ è detto Score $S(x;\theta)$.
Lo score misura come varia la probabilità associata al campionamento quando varia il parametro θ. Risulta comodo riscrivere lo score come

$$S(x,\theta) = \partial_\theta \ln\left[pdf_x(x;\theta)\right].$$

Lo score è una funzione della random variable x, quindi il suo valore fluttua statisticamente e il singolo valore non è particolarmente significativo mentre lo sono $E[S(x;\theta)]$ e $E[S^2(x;\theta)]$ che non dipendono da x ma solo da θ.
I valori di aspettazione sono definita da

$$E[S(x;\theta)] = \int S(x;\theta)\, pdf_x(x;\theta)\, \mathrm{d}x = f(\theta)$$

$$E[S^2(x;\theta)] = \int S(x;\theta)^2\, pdf_x(x;\theta)\, \mathrm{d}x = g(\theta)$$

Se vale

- l'intervallo di integrazione (il sample space Ω su cui è definita la *pdf*), non dipende da θ;
- la pdf_x è una funzione regolare;

allora si dimostra che $E[S(x;\theta)] = 0$, cioè il valor medio dello score è nullo, che è il motivo per cui si fa riferimento al valore quadratico medio.

Score per N campionamenti. Per un set di campionamenti IID $\{x_1, \ldots, x_N\}$ di una $pdf_x(x;\theta)$ si definisce lo score come soma degli score dei singoli campionamenti:

$$\begin{aligned} S(\vec{x};\theta) &= \sum_{i=1}^{N} S(x_i;\theta) = \sum_{i=1}^{N} \partial_\theta \ln\left[pdf_x(x_i;\theta)\right] = \partial_\theta \sum_{i=1}^{N} \ln[pdf_x(x_i;\theta)] \\ &= \partial_\theta \ln\left[\prod_{i=1}^{N} pdf_x(x_i;\theta)\right] = \partial_\theta \ln \mathcal{L}(\vec{x};\theta). \end{aligned}$$

questa è la relazione tra la likelihood e lo score. Tale relazione soddisfa le condizioni precedenti.

Informazione di Fischer. Si definisce l'informazione contenuta nel set IID $\{x_1, \ldots, x_N\}$ e relativa al parametro θ il valore quadratico medio dello score:

$$I_{\vec{x}}(\theta) = E[S(\vec{x};\theta)^2].$$

quando valgono le condizioni precedenti (e quindi il valore medio dello score è nullo) si può scrivere:

$$I_{\vec{x}}(\theta) = E\left[-\partial_\theta^2 \ln \mathcal{L}(\vec{x};\theta)\right] = E\left[-\partial_\theta S(\vec{x};\theta)\right].$$

L'informazione è un numero, quindi le espressioni precedenti sono da valutare in corrispondenza del valore vero del parametro.
Si dimostra che $I_{\vec{x}}(\theta)$ soddisfa i requisiti che si sono posti per una metrica che misuri l'informazione contenuta nei dati campionati e relativa ad un parametro. Rao-Cramer lega l'informazione alla varianza dello stimatore:

$$\mathrm{Var}\left[\hat{\theta}\right] \geq \frac{(1+\partial_\theta b_N[\hat{\theta}])^2}{E[I_{\vec{x}}(\theta)|_{\theta=\theta_V}]} = \frac{(1+\partial_\theta b_N[\hat{\theta}])^2}{E[-\partial_\theta^2 \ln \mathcal{L}(\vec{x};\theta)|_{\theta=\theta_V}]}.$$

e quindi spiega come maggiore è l'informazione contenuta nei dati e minore è la varianza dello stimatore. Lo stimatore è efficiente se vale il segno di uguale.

Per riassumere:

- la likelihood è la joint-*pdf* dei campionamenti. Questa è una funzione dei parametri incogniti quando si fissa l'insieme dei dati.
- L'informazione di Fischer è il valore medio della derivata seconda della likelihood; se la likelihood è una campana, allora l'informazione dipende da quanto tale campana è stretta.
- Si è introdotto un metodo per costruire uno stimatore usando la likelihood: la maximum likelihood.

Metodo della massima verosimiglianza. La likelihood può essere usata per costruire uno stimatore del parametro θ secondo la procedura che segue:

- si studia la likelihood come funzione di θ.
- si cerca il valore per cui $\mathcal{L}(\theta)$ oppure $\ln \mathcal{L}(\theta)$ assume un massimo assoluto (sul dominio in cui è definita). Tale massimo è $\hat{\theta}_{\mathrm{ML}}$.

Invarianza per trasformazione. Si consideri una $pdf(x;\theta)$ e la likelihood $\mathcal{L}(\vec{x};\theta)$ corrispondente ad N campionamenti. La trasformazione $\lambda = \lambda(\theta)$ porta a $pdf(x;\lambda(\theta))$ e quindi ad una $\mathcal{L}(\vec{x};\lambda(\theta))$. Poiché vale la relazione

$$\partial_\theta \mathcal{L}(\vec{x};\lambda(\theta)) =_\lambda \mathcal{L}(\vec{x};\lambda(\theta))\,\partial_\theta \lambda.$$

si ricava la relazione tra gli stimatore di ML per λ e θ:

$$\hat{\lambda}_{\mathrm{ML}} = \lambda(\hat{\theta}_{\mathrm{ML}}).$$

Lo stimatore della ML può avere bias non nullo anche se è possibile dimostrare che, sotto ipotesi molto generiche, è asintoticamente unbiased. Similmente si dimostra che gli stimatori ML sono asintoticamente efficienti. Inoltre, si dimostra che se esiste uno stimatore efficiente, allora lo stimatore ML è tale. L'invarianza per trasformazioni preserva il punto di massimo, sia l'ordinata che l'ascissa; però non preserva le stesse proprietà dello stimatore.

Stimatori della maximum likelihood. Per $N \to \infty$ si ha

- $\hat{\theta}_{\mathrm{ML}}$ è unbiased ed efficiente
- $I(\theta) = I(\hat{\theta}_{\mathrm{ML}})$
- il MVB si traduce in
$$\sigma^2_{\hat{\theta}_{\mathrm{ML}}} = \mathrm{Var}\left[\hat{\theta}_{\mathrm{ML}}\right] = \frac{1}{\sigma^2}.$$
- l'espressione analitica della likelihood è quella di una gaussiana, quindi $\ln \mathcal{L}$ è una parabola
$$\mathcal{L}(\theta) = \mathcal{L}_{\max} e^{-\frac{(\theta-\hat{\theta}_{\mathrm{ML}})^2}{2I(\hat{\theta}_{\mathrm{ML}})}}.$$

In generale si afferma che

- le proprietà elencate sono spesso valide anche nel caso di N piccolo, se lo stimatore è unbiased ed efficiente
- se esiste uno stimatore efficiente allora $\hat{\theta}_{\mathrm{ML}}$ lo è
- vale la proprietà di invarianza per trasformazione del parametro, se $\alpha = \alpha(\theta)$:
$$\hat{\alpha}_{\mathrm{ML}} = \alpha(\hat{\theta}_{\mathrm{ML}}) \implies \mathcal{L}_{\max} = \mathcal{L}(\hat{\alpha}_{\mathrm{ML}}) = \mathcal{L}(\hat{\theta}_{\mathrm{ML}}).$$

Lecture 10

gio 16 dic 2021 13:30

Si discutono i metodi di stima della varianza. Lo stimatore della maximum likelihood è una funzione dei campionamenti. La funzione likelihood risulta essere

$$\mathcal{L}(x_1,\ldots,x_N;\theta) = \prod_{i=1}^{N} pdf(x;\theta).$$

lo stimatore è quello per cui $\mathcal{L}$ ammette un massimo assoluto. Facendo una stima si vuole stabilire la bontà di uno stimatore.
Asintoticamente lo stimatore ha tutte le caratteristiche desiderate:

1 unbiased

2 efficiente

3 l'informazione rispetto al parametro si può stimare con l'informazione per $\theta_V \approx \hat{\theta}_M L$

4 il MVB risulta essere
$$\mathrm{Var}\left[\hat{\theta}_{\mathrm{ML}}\right] = \frac{1}{I(\hat{\theta}_{\mathrm{ML}})}.$$

5 se λ è un nuovo modo per definire il parametro θ: $\lambda = \lambda(\theta)$ si ha invarianza per trasformazione dei parametri:
$$\hat{\lambda}_{\mathrm{ML}} = \lambda(\hat{\theta}_{\mathrm{ML}}).$$

6 assume la forma di una gaussiana con parametri σ^2 da Rao-Cramer e $\mu = \hat{\theta}_{\mathrm{ML}}$. Tuttavia, la gaussiana è una pdf, mentre la likelihood, nell'approccio bayesiano, è la distribuzione di probabilità di θ (cioè la probabilità di arrivare vicino al valore vero); ma ciò non è vero per l'approccio frequentista. Con bayes si associa [r]

ad esempio $pdf(x;\theta) = Ke^{-\frac{x}{\theta}} = He^{-\lambda x}$, dove K e H dipendono dal fatto che l'integrale sul dominio dev'essere uno; si è posto $\lambda = \lambda(\theta) = \frac{1}{\theta}$. Infatti, $\partial_\lambda \mathcal{L} = \partial_\theta \mathcal{L}\, \partial_\lambda \theta$.

Idealmente si può arrivare a tali condizioni soddisfatte, ma nei casi concreti non lo sono necessariamente.
Spesso, per N grande, si può assumere di lavorare in tali condizioni: in particolare che lo stimatore sia privo di bias ed efficiente per cui vale Rao-Cramer per determinare la sua varianza. Se esiste uno stimatore efficiente, allora quello della likelihood è efficiente.
La tecnica di cercare il massimo della likelihood non dice come trovare la varianza dello stimatore. Si hanno tre metodi:

- diretto, fa uso della quarta proprietà, si assume che valga Rao-Cramer, che lo stimatore sia unbiased ed efficiente. Da cui segue
$$\mathrm{Var}\left[\hat{\theta}_{\mathrm{ML}}\right] = \frac{1}{I[\hat{\theta}_{\mathrm{ML}}]} = \frac{1}{-\partial_\theta \ln \mathcal{L}\big|_{\theta=\hat{\theta}_{\mathrm{ML}}}}.$$
[r] Generalmente, esso non è utilizzato.
- grafico; questo metodo, come il successivo, si può applicare anche quando lo stimatore non è unbiased ed efficiente. Spesso tali due tecniche si usano simultaneamente e consentono anche di esplorare un particolare bias.
- Monte Carlo;

Metodo grafico. Si consideri il logaritmo della likelihood che è funzione solo del parametro quando si fissano i campionamenti. Si isola una funzione vicino al massimo:
$$\ln \mathcal{L}(\theta) = \ln \mathcal{L}(\hat{\theta}_{\mathrm{ML}}) + \partial_\theta \mathcal{L}(\theta - \hat{\theta}_{\mathrm{ML}}) + \frac{1}{2}\partial_\theta^2 \mathcal{L}(\theta - \hat{\theta}_{\mathrm{ML}})^2.$$
uno sviluppo al secondo ordine consiste in approssimare con una parabola. Il punto $\hat{\theta}_{\mathrm{ML}}$ è un massimo, dunque la derivata prima è nulla. Mentre
$$\partial_\theta^2 \mathcal{L} = -\frac{1}{\sigma^2}.$$
vale la seconda uguaglianza quando si suppone l'uguale in Rao-Cramer e non ci sia bias. Pertanto
$$\ln \mathcal{L}(\theta) = \ln \mathcal{L}(\hat{\theta}_{\mathrm{ML}}) = -\frac{1}{2}\frac{1}{\sigma^2}(\theta - \hat{\theta}_{\mathrm{ML}})^2.$$
ponendo $\theta = \hat{\theta}_{\mathrm{ML}} \pm \sigma$ si ha
$$\ln \mathcal{L}(\hat{\theta}_{\mathrm{ML}} \pm \sigma) = \ln \mathcal{L}(\hat{\theta}_{\mathrm{ML}}) - \frac{1}{2}.$$
questo vale per uno stimatore efficiente e privo di bias; ma si ha solamente in un caso asintotico. Tuttavia, la relazione 5 afferma che

Sebbene una likelihood finita non sia una likelihood asintotica, comunque dato che gli intervalli della varianza si corrispondono, allora si può mantenere lo stesso concetto di varianza anche per una likelihood finita.
Più una likelihood assomiglia una parabola, segue che l'approssimazione tramite il metodo grafico risulta essere migliore.

Toy Monte Carlo. Questo metodo è spesso utilizzato in concomitanza al metodo grafico. Tramite una simulazione numerica si ripete più volte un esperimento. Si assume di conoscere la *pdf* che descrive la misura; bisogna capire la sua forma in basse alle misure stesse. Una volta ottenuta la *pdf* si possono simulare i campionamenti a cui si applica la stessa tecnica di analisi utilizzata per i dati veri, originali. Da ogni set di dati si estrae un $\hat{\theta}_{\mathrm{ML}}$ e pure la varianza tramite il metodo grafico. Con abbastanza toy esperimenti si può anche rappresentare la distribuzione dello stimatore.
Per fare le simulazioni bisogna fissare il valore del parametro. Ci si aspetta che la media sia pari al valore del parametro. Qualora non fosse così, si conclude che la tecnica adottata è una procedura che induce un bias. Potrebbe anche darsi che la *pdf* scelta non è quella che rappresenta l'esperimento.
A tal punto la larghezza della curva dà informazioni sulla varianza. Se tale curva è una gaussiana, si può eseguire un fit, tramite cui si può stabilire un intervallo di confidenza $[\hat{\theta}_{\mathrm{ML}} \pm \sigma^{\mathrm{toy}}]$ del 68%. Essendo frequentista, si afferma di avere una probabilità del 68% che l'intervallo contenga il valor vero (e non il viceversa, proprio perché si è frequentisti).

Nella fisica delle particelle, quando si dichiara una scoperta, bisogna che l'evento inatteso dev'essere a più di 5σ; cioè avere una discrepanza maggiore di 5σ dal valor atteso. A 3σ si ha un'indicazione.

[r] Si supponga di compiere una misura di una variabile $y = d\sin\theta$, con $\theta_1, \ldots, \theta_N$, $d = 10 \pm 3$ misurato una singola volta. Per ogni θ si propaga l'errore su y sia da θ che da d. Tuttavia, l'errore di d non è una fluttuazione statistica perché si è misurato una singola volte.
La procedura corretta è calcolare $\langle\theta\rangle$ [r]

7.3.1 Extended Maximum Likelihood

Il numero di campionamenti N può essere ignoto. Il caso è tipico di fisica delle particelle. Pertanto, si aggiunge un fattore correttivo alla maximum likelihood:

$$\mathcal{L}(x_1, \ldots, x_N; \theta) = \frac{e^{-\nu}\nu^N}{N!} \prod pdf(x_i; \theta).$$

7.4 Minimi quadrati

Il metodo dei minimi quadrati è utile per stimare i valori dei parametri a partire da dati sperimentali. Esso può anche rendersi utile per capire quale sia il modello migliore che descriva tali dati.
Si supponga che i dati di una variabile abbiano incertezza trascurabile rispetto quella dell'altra variabile. Ci sono casi in cui l'incertezza è nulla ed altri in cui l'incertezza è da stimare. [r]

la miglior stima dei parametri risulta essere

$$Q^2 = \sum \frac{(y - f(x;\theta))^2}{\sigma^2}.$$

si scelgono i parametri tali per cui Q sia minimo. Questo si può fare anche se i dati non sono distribuiti in modo gaussiano. [r]

Lecture 11

gio 23 dic 2021 13:30

Si considera un problema più generale di quello affrontano con la maximum likelihood: si ha un modello che definisce la relazione tra due grandezze fisiche X e Y che dipendono da k parametri:

$$Y = f(\vec{\theta}, X), \quad \vec{\theta} = (\theta_1, \dots, \theta_k)$$

Si stimano i valori dei parametri a partire da dati sperimentali: N misura y_i del valore della grandezza Y effettuate in corrispondenza di un valore x_i della grandezza X.
Le coppie (x_i, y_i) sono estrazioni da una *pdf* e vi si associano le rispettive deviazioni standard σ_{x_i} e σ_{y_i}. I dati sono tipicamente rappresentati in un grafico con barre di errore pari alle larghezze delle *pdf*. Si trova un metodo generale per stimare i parametri $\theta_1, \dots, \theta_k$ a partire dai dati. Si parla di interpolazione o fit dei dati.
Il metodo dei Mini Quadrati è quello di scegliere come stima di $\vec{\theta}$ il valore per cui il seguente funzionale è minimo

$$Q^2(\theta) = \sum_{i=1}^{N} \frac{\left[y_i - f(\vec{\theta}, x_i)\right]^2}{\sigma_i^2}$$

tale funzionale è la somma dei quadrati delle distanze tra i punti campionati e la funzione $Y = f(\vec{\theta}, X)$. Ciascuna distanza è normalizza alla larghezza della $pdf(y_i)$ e quindi tiene conto delle fluttuazioni statistiche caratteristiche del campionamento. Si cerca il valore di $\vec{\theta}$ per cui Q^2 ammette minimo assoluto nello spazio dei parametri:

$$\partial_{\theta_j} Q^2 = 0, \quad j = 1, \dots, k$$

Si cerca il modello sia il più vicino ad i dati. Per errori distribuiti in modo gaussiano, i minimi quadrati coincidono con la maximum likelihood.

Osservazione. Sebbene pure le x_i hanno varianza, [r] si può ricondursi alla situazione in cui solamente le y_i hanno varianza. Infatti, basta propagare l'errore di x_i su $f(x, \vec{\theta})$ e quindi su $y_i - f(x, \vec{\theta})$.[r] riportare sulle y la grandezza che presenta gli errori maggiori.

Osservazione. La posizione del minimo non dipende dai σ_i perché infatti si può scrivere

$$Q^2 = \frac{1}{\sigma^2} \sum \frac{(y_i - f(x, \vec{\theta}))^2}{\frac{\sigma_i^2}{\sigma^2}}$$

dove σ^2 è la media delle varianze. Dunque, quanto importa sono le incertezze relative come suggerito dal denominatore. Quindi il valore del minimo non dipende dall'incertezza [r] che invece influenza le incertezze dei parametri.

Caso gaussiano. Se la *pdf* è gaussiana allora gli stimatori coindicono. Se θ_{LS} è unbiased ed efficiente allora MVB garantisce che

$$\text{Var}\left[\theta_{\text{LS}}\right] = \text{Var}\left[\theta_{\text{ML}}\right] = -\frac{1}{E[\partial_\theta^2 \mathcal{L}]_{\theta=\theta_{\text{LS}}}}$$

Si consideri il caso con x_i senza errore, mentre y_i hanno σ_i come incertezza. La distribuzione di ciascun y_i è gaussiana:

$$pdf(y_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(y_i - \mu_i)^2}{2\sigma_i^2}}$$

il modello dei minimi quadrati afferma che $\mu_i = f(x_i, \vec{\theta})$. Per la likelihood si hanno variabili IID. Si consideri la variabile ausiliaria, intesa come trasformazione

$$z_i = \frac{y_i - f(x_i, \vec{\theta})}{\sigma_i}$$

il valore x_i non è una variabile casuale, l'unica che lo è risulta essere y_i. Questa variabile è distribuita secondo una gaussiana standard. [r] La likelihood è il prodotto delle *pdf*

$$\mathcal{L}(y_1, \dots, y_N) = \prod \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(\dots)^2}{2\sigma_i^2}} \implies \ln \mathcal{L} = \ln \text{cost.} - \frac{1}{2} \underbrace{\sum \frac{(y_i - f(x_i, \vec{\theta})}{\sigma_i^2}}_{Q^2}$$

Quando si cerca il massimo della likelihood, le derivate si annullano nello stesso punto. Si calcola l'incertezza associata ai parametri. Si è già visto lo sviluppo attorno al massimo della likelihood:

$$\ln \mathcal{L}(\theta) = \ln \mathcal{L}(\theta_{\text{ML}}) - \frac{1}{2} \frac{(\theta - \theta_{\text{ML}})^2}{\sigma^2}$$

per i minimi quadrati, lo sviluppo diventa

$$Q^2(\theta) = Q^2(\theta_{\text{LS}}) + \frac{(\theta - \theta_{\text{LS}})^2}{\sigma^2}$$

[r] dunque si sale di una quantità pari ad uno sul grafico e si ottiene la larghezza della distribuzione dei parametri.

7.5 Test del chi quadro

Si introduce un test di ipotesi per misurare quanto il modello sia compatibile con i dati raccolti. [r] Nel metodo dei minimi quadrati si ha

$$f(x, \vec{\theta}_{\text{V}}) = E[y_i] = \mu_i, \quad f(x, \vec{\theta}_{\text{LS}}) \approx E[y_i]$$

dunque risulta

$$Q^2 = \sum \frac{(y_i - f(x, \vec{\theta}))^2}{\sigma_i^2} = \sum \frac{(y_i - \mu_i)^2}{\sigma_i^2}$$

inoltre

$$Q^2(\theta_{\text{V}}) = \sum \frac{(y_i - \mu_i)^2}{\sigma_i^2} \equiv \chi^2$$

è una variabile casuale perché somma di variabili casuali. Tale chi quadro ha $N - k$ gradi di libertà, dove k è il numero di parametri che si determinano dai dati. Pertanto

$$Q^2(\theta_{\text{LS}}) \approx \chi^2$$

il valore assunto sul minimo di Q^2 è χ^2 che a sua volta è una variabile casuale che segue la distribuzione omonima. Il valore di aspettazione è $N - k$ cioè i gradi di libertà. Per ogni insieme di misure con cui si fa un fit, si ha un χ^2 diverso.
Da un solo insieme di misure si possono dedurre varie implicazioni. La prima è quella di aspettarsi un chi quadro ridotto prossimo ad uno. Risulta evidente che una misura del χ^2 nelle due code è un risultato improbabile ed esso non è un valore bello. Si potrebbe aver sbagliato a raccogliere i dati. Oppure il chi quadro porta un messaggio differente.
Un valore prossimo allo zero riflette il fatto che il modello passa molto vicino ai dati. Questo può succedere quando si sbaglia a valutare l'errore oppure si abbiano tanti parametri liberi quanti sono i punti raccolti. Si ha un grande raccordo con il modello.
Per valori molto maggiori del valore di aspettazione indica che gli scarti dal modello sono molto grandi. Questo può succedere quando il modello con cui si vuole descrivere i dati potrebbe non essere quello corretto. Si ha un piccolo raccordo con il modello.
La scelta delle regioni è arbitraria, tipicamente si utilizza una p-value di 5%:

$$p = \int_{+\infty}^{\chi^2_{\text{misur}}} pdf(\chi^2)$$

se p è sopra il 5% allora si è nell'area accettabile.
Si sta facendo un test di ipotesi che testa se i dati sono statisticamente compatibili; non si confrontano modelli. Se il risultato sotto il 5% si è dimostrato che il modello non vale. Se il risultato sta sopra, non si è dimostrato che vale, ma solo che potrebbe valere. Potrebbero esistere modelli più compatibili con i dati. Confronta Popper e la falsificabilità. [r]
Tipicamente si costruisce una statistica t e si osserva la sua distribuzione in base a quale modello si sceglie. In base a dove si misura il valore della statistica, segue che si può escludere un modello o l'altro. Tuttavia, il problema si presenta quanto la misura di t cade in una regione comune ad entrambe le distribuzioni: bisogna scegliere un compromesso con cui accettare un certo numero di falsi negativi e falsi positivi.
Il caso del χ^2 è speciale perché non si confrontano ipotesi e si ha una sola regione di disaccordo con il modello. Esso, come molti test di ipotesi, dà un risultato sull'insieme di tutti i dati che si sono raccolti e ciò non è ideale, perché potrebbe non evidenziare problemi che si trovano nei dati.
Un modo utile per studiare quando un modello è buono è quello di rappresentare i residui normalizzati detti pull:

$$\frac{y_i - f(x_i)}{\sigma_i}$$

ci si aspetta che siano distribuiti in modo casuale attorno allo zero. Con tanti dati ci si aspetta una distribuzione gaussiana standard. Se la distribuzione non appare casuale, allora significa che il modello può (anche in parte) non descrivere bene i dati. Se si ha una residua dipendenza allora c'è un problema nel modello. Tale metodo risulta molto potente ed è sempre disponibile. Dunque, bisogna capire se cambiare modello, se il modello ha dei limiti di validità, etc.

Bisogna stare attenti alla sovra-correlazione tra i parametri: $y = ab + cx$ ha infinite soluzioni per a e b, tutte quelle che hanno lo stesso prodotto. Bisogna guardare attentamente la matrice di covarianza.

Interpolazione ed estrapolazione. Spesso si desidera valutare il modello in un punto differente da quelli sperimentalmente misurati. Il valore medio si trova per sostituzione. Tuttavia, per calcolare l'incertezza bisogna considerare la matrice di covarianza. Si deve propagare l'incertezza e la covarianza.
Inoltre, più ci si allontana dai dati più l'incertezza dell'estrapolazione è maggiore.

Teorema di Gauss-Markov. Per un modello lineare nei parametri, la soluzione dei minimi quadrati si trova per via analitica; sia del minimo che per la varianza. In ROOT è TLinearfitter. Vale il teorema di Gauss-Markov: se il modello è lineare nei parametri, allora lo stimatore dei minimi quadrati è privo di bias e tra tutti i possibili stimatori lineari nei parametri, quello dei minimi quadrati è quello con varianza minore.