COMPLEMENTI DI STATISTICA

Richiami di carattere generale

Fulvio Ricci

Dipartimento di Fisica, Università di Roma *La Sapienza*

INDICE

Introduzione.

Richiamo di alcuni concetti fondamentali di probabilità.

Distribuzioni discrete di probabilità: la distribuzione binomiale e la distribuzione di Poisson.

Distribuzioni continue di probabilità: la distribuzione di Gauss.

INTRODUZIONE.

Per valutare l' affidabilità di una misura occorre ripeterla parecchie volte ed esaminare i diversi risultati ottenuti. In questo modo, avendo la possibilitá (almeno concettuale) di ripetere l'esperimento sempre sotto le stesse condizioni, si puó fare uso dei metodi statistici per trarre conclusioni sullo stato della grandezza oggetto di studio ed in particolare sul suo *valore vero*. Va fatto presente che non tutti i tipi di incertezze sperimentali sono valutabili da un'analisi statistica basata su misure ripetute. Infatti proprio questa osservazione è alla base della distinzione tra errori **casuali** ed errori **sistematici**. L'esempio tipico è quello dell'errore compiuto nella misura di un intervallo di tempo per mezzo di un cronometro. Si consideri ad esempio la misura del periodo d'oscillazione di un pendolo.

Una sorgente d'errore è determinata dal tempo di reazione dell'operatore nel far partire e nel fermare il cronometro. Dal momento che, ripetendo piú volte la misura, l'operatore a volte ritarderá l'istante di partenza (o di arresto) e altre volte lo anticiperá, e ció avverrá con uguale probabilitá, potremo trattare questo tipo di errore come un classico errore casuale. Analizzando quindi i risultati ed in particolare come differiscano l'uno dall'altro, possiamo ottenere una "buona" stima statistica di questo genere di errore.

D'altra parte, se il nostro cronometro marcia *costantemente* lento, allora tutti i tempi sono sottostimati e la ripetizione delle misure fatta con lo stesso cronometro

non rivela questo genere d'errore, l'errore **sistematico**.

La distinzione tra errore sistematico ed errore casuale non è sempre netta. Ad esempio l'errore di lettura dovuto ad un effetto di parallasse potrebbe essere sistematico se l'operatore effettuasse tutte le misure guardando l'ago dello strumento sempre con un angolo d'inclinazione positivo (o sempre negativo).

Non vi è una semplice teoria che ci dica cosa dobbiamo fare circa gli errori sistematici. In realtá l'unica considerazione generale circa questi ultimi è che essi debbano essere individuati e ridotti fino a diventare molto minori della precisione richiesta. Laddove ció non sia possibile occorrerá valutare come vadano combinate le incertezze dovute agli errori sistematici e quelle dovute alle componenti casuali. Supponiamo allora che tramite un'analisi statistica di vari risultati sia possibile fornire la componente casuale dell'errore $\delta G_{casuale}$ sulla grandezza G oggetto della misura. Inoltre, se la grandezza G è frutto di una misura indiretta, sará possibile dedurre l'errore sistematico $\delta G_{sistematico}$ derivandolo dall'errore di sensibilitá degli strumenti utilizzati.

Una volta stimato $\delta G_{sistematico}$ potremo procedere a calcolare δG_{tot}. Nell'ipotesi che i due contributi all'errore totale siano indipendenti, l'errore complessivo sará ottenuto sommando quadraticamente la componenete sistematica a quella casuale:

$$\delta G_{tot}{}^2 = \delta G_{casuale}{}^2 + \delta G_{sistematico}{}^2$$

Sommare quadraticamente gli errori è un modo per tener conto dell'ipotesi di perfetta indipendenza dei due

contributi portandoci a credere che ció dovrebbe dare luogo a qualche possibile cancellazione di incertezze. Occorre comunque chiarire che questo modo di procedere è empirico e non deve indurre nell'errore di interpretare ${\delta G_{tot}}^2$ come la varianza totale della grandezza G, poiché questo sottointenderebbe che i risultati delle misure siano tutti affetti da errori statistici indipendenti. Tuttavia questa espressione quadratica fornisce una stima ragionevole (anche se non rigorosamente vera) dell'incertezza totale della misura.

Va infine notato che proprio gli errori sistematici che non si riducono aumentando il numero di misure definiscono il limite ultimo della qualitá dell'apparato sperimentale utilizzato.

RICHIAMI DI ALCUNI CONCETTI FONDAMENTALI DI PROBABILITÀ.

Un concetto basilare della statistica è quello della *distribuzione di frequenza* per i risultati di un determinato esperimento. Questa distribuzione è una successione discreta o una funzione continua che descrive la frazione di volte in cui si presenti ciascuno dei vari possibili risultati di una misura ripetuta N volte.

Si definisce *distribuzione di probabilitá* la distribuzione di frequenza ottenuta al limite per N, numero di prove, che tende all'infinito.

Determinare quindi la distribuzione di frequenza per i risultati di un esperimento rappresenta tipicamente il primo approccio al problema di analizzare statisticamente i dati. È ovvio che occorra stabilire quale sia la qualitá della descrizione statistica ottenibile dai dati disponibili. Infatti, avere l'informazione completa sull'esperimento significherebbe essere in possesso di tutte le osservazioni possibili relative all'esperimento in corso. Nel linguaggio statistico, mutuato dagli studi demografici, questo corrisponderebbe a conoscere tutta la *popolazione*, cioe' l'insieme di tutti i possibili risultati; la conoscenza completa della popolazione non è generalmente disponibile allo sperimentatore che al contrario avra' a sua disposizione, a seguito dell'esperimento; solo una frazione di tutti i possibili risultati, cioé un *campione della popolazione.* Ad esempio, ripetendo N volte la misura del periodo di un pendolo T avremo un campione di N elementi della popolazione di dimensione infinita

che rappresenta lo *spazio dei possibili risultati.*
Dai risultati delle misure, partendo cioé dal campione di dimensione N, possiamo calcolare il valor medio $\bar{t}$, il cui significato credo sia palese. Comunque piú tardi vedremo che possa essere stimato il valor medio sulla base della nostra conoscenza limitata al solo campione della popolazione. Nel caso in cui ci si riferisca *all'intera popolazione* allora tale valor medio lo indicheremo con μ o in modo equivalente con il simbolo $E[\bar{T}]$

$$\mu = E[\bar{T}]$$

è il *valore aspettato* della popolazione dei dati.
Con il simbolo

$$\bar{t}$$

indichiamo la media eseguita sul campione dei dati e quindi rappresenta una stima del valore aspettato.
In un linguaggio piú consono a quello di uno sperimentatore, diremo che $\bar{t}$, la media di quel gruppo di dati, rappresenta una stima del *valore vero* $E[T]$.
Abbiamo detto che la media di un campione è definibile sulla base della conoscenza del campione, o meglio della distribuzione di frequenza associata al campione stesso (la distribuzione campionaria).
Innanzitutto, il campione dei risultati puó essere analizzato raggruppando in vari dati in modo da costruire la distribuzione campionaria. Per fare questo, l'intervallo in cui cadono tutti i risultati viene diviso in sottointervalli contigui e non sovrapposti *classi.* Quindi conteremo il numero di volte n_i che i dati del campione cadono in

ciascun intervallo. Riferendoci all'esempio della misura del periodo del pendolo, supponiamo di avere un campione di N=10 misure effettuate con un cronometro che misura in secondi il periodo del pendolo. Sia ad esempio il campione costituito dai seguenti dati espressi in secondi (s)

26.4, 23.9, 25.1, 24.6, 22.7, 23.8, 25.1, 23.9, 25.3, 25.4

Se suddividiamo in sottointervalli di 1 s lo spazio dei risultati, otteremo

tra *22* - *23* s n_1 =1 risultato
tra *23* - *24* s, n_2 =3 risultati
tra *24* - *25* s, n_3 =1 risultato
tra *25* - *26* s , n_4 =4 risultati
tra *26* - *27* s, n_5 =1 risultato
tra *27* - *28* s, n_6 =0 risultati

In questo modo posssiamo definire la media della distribuzione come

$$\bar{T} = \frac{\sum_i n_i t_i}{N}$$

dove t_i è il valore centrale dell'i-esimo intervallo in cui ho suddiviso lo spazio dei risultati.
Le quantitá

$$f_i = \frac{n_i}{N}$$

definiscono la distribuzione campionaria delle frequenze f_i.

Puó essere dimostrato che in generale vale la relazione

$$\sum_{i=1}^{N} f_i = 1$$

come é immediatamente verificabile dall'esempio precedente.

La media della popolazione è anche detta media pesata e $\bar{t}$ rappresenta proprio la somma dei vari valori t_i moltiplicati per l'*occorrenza* , cioe' la frazione f_i dell'iesimo risultato.
Aumentando il numero di misure e, contemporaneamente infittendo gli intervalli di divisione della porzione di spazio che stiamo considerando, la distribuzione di frequenza tende ad una funzione f della variabile casuale T che chiameremo funzione di distribuzione di probabilitá.
La T rappresenta il possibile risultato dell'esperimento e puó rappresentare altresí una funzione, definita sempre nello spazio dei risultati, il cui valore dipende da un sol punto (detto anche evento) dello spazio dei risultati, e non è condizionata da alcuna altro evento dello spazio.
Questo spazio dei risultati (o degli eventi) puó essere numerabile (discreto). L' esempio classico di spazio degli eventi discreto è quello relativo allo spazio dei possibili riultati ottenibili lanciando una moneta o un dado.

Se $x = x_i$ rappresenta il numero di volte in cui si ottiene "testa" lanciando la moneta piú volte, è chiaro che alla variabile discreta "*numero di volte* per cui si ottiene testa" sará associata una funzione di distribuzione discreta.

Essa comunque sará tale che

$$0 \leq f(x_i) \leq 1$$

$$\sum_{i=1}^{\infty} f(x_i) = 1 \qquad con \qquad x \in \{x_i\}$$

Si definisce accanto alla funzione di distribuzione f la funzione *cumulativa di distribuzione* come

$$F(x_0) = \Sigma_{x_i \leq x_0} f(x_i)$$

Se, ad esempio, si desidera conoscere la probabilità che x sia compresa nell'intervallo $a \leq x \leq b$ allora utilizzando la funzione cumulativa avremo

$$P(a \leq x \leq b) = \sum_{a \leq x_0 \leq b} f(x_i) = F(b) - F(a)$$

Se **X** è una variabile casuale (detta anche variabile *aleatoria*) continua la distribuzione limite sará una funzione continua $f(x)$ ed il prodotto $f(x)dx$ rappresenterà la probabilità che un risultato cada in $(x, x + dx)$.
Posso allora definire per analogia la funzione cumulativa di probabilità come la probabilità che un risultato sia compreso in (y, a). Nel caso continuo la sommatoria è sostituita dall'integrale della forma differenziale $f(x)dx$ cosí che avremo:

$$F(y) - F(a) = \int_a^y f(x)dx$$

La $f(x)$ è anche detta funzione densità di probabilitá: infatti, se $f(x)$ assume valori piú grandi in una certa zona dello spazio degli eventi, quando effettueremo le

misure, i risultati andranno ad addensarsi piú in questa regione che non nelle altre zone ove $f(x)$ assume valori piú bassi.

Per introdurre questi concetti base quali il valore aspettato e le funzione di distribuzione abbiamo implicitamente seguito quella scuola di probabilisti che introducono il concetto fondamentale di probabilità come limite per il numero di tentativi (o casi possibili) che tendono all'infinito. Questa scelta non è stata dettata da una profonda convinzione, ma solo dal desiderio di introdurre questi concetti fondamentali in modo rapido ed intuitivo.

DISTRIBUZIONI DISCRETE DI PROBABILITÀ: LA DISTRIBUZIONE BINOMIALE E LA DISTRIBUZIONE DI POISSON.

La distribuzione di probabilità piú semplice è la distribuzione binomiale. Essa è relativa a tutti quei processi che siano riconducibili al dilemma "successo" o "insuccesso".
Questo è un tipo di punto di vista assumibile in quasi tutti i processi osservativi. Si pensi ad esempio al caso in cui si stia effettuando il monitoraggio continuo di una grandezza le cui fluttuazioni siano influenzate da variazione di molteplici parametri. Potrebbe essere il caso della misura ripetuta periodicamente nel tempo (monitoraggio) della tensione d'uscita di un sismografo. Se l'interesse fondamentale quello di stabile se ad un certo istante siamo in presenza di un segnale sovrapposto alle fluttuazioni casuali, potremo assumere il punto di vista binomiale che consiste nel limitarsi ad osservare quando il segnale di monitor ha superato un valore di soglia da noi prefissato e battezzeremo questo evento (la rivelazione di un sisma) come un successo nel senso della distribuzione binomiale. Se indichiamo con p la probabilità di osservare un successo e q quella dell'insuccesso (ovviamente $p + q = 1$), é facile rendersi conto che la probabilità di osservare ν successi in N osservazioni sará espressa

$$P_{N,p}(\nu) = \binom{N}{\nu} p^{\nu} q^{N-\nu}$$

dove abbiamo fatto uso della proprietá che la probabilità di accadimento di piú eventi indipendenti è il prodotto della probabilità di ciascun evento.

L'esempio classico che si porta nel presentare la distribuzione binomiale è quello del lancio di una moneta effettuato piú volte. Si puó ad esempio stabilire che il successo è caratterizzato dal fatto che lanciando una moneta compaia il segno "croce". Sia $p = 1/2$ e per conseguenza $q = 1 - p = 1/2$. Possiamo quindi calcolare in $N = 10$ prove quale sia la probabilità di ottenere $n = 5$ volte "croce". Essa sará data da

$$P_{n,1/2} = \binom{N}{\nu} p^{\nu} q^{N-\nu} = \binom{10}{5} (1/5)^1 (1/2)^5$$

Infine possiamo calcolare il valore medio della distribuzione (detto anche valore *aspettato* $E[n]$) cioé il numero medio di successi su N prove

$$E[n] = \sum_{n=0}^{N} n P_{N,p}(n) = \sum_{n=0}^{N} n \binom{N}{\nu} p^{\nu} (1-p)^{N-\nu}$$

Sviluppando questa definizione si ottiene che

$$E[n] = Np$$

L'importanza di questa distribuzione risiede nel fatto che qualsivoglia esperimento puó essere ricondotto ed interpretato alla luce della distribuzione binomiale. Infatti possiamo sempre identificare tra tutti i possibili risultati di un esperimento un particolare risultato come "favorevole" o di "successo" e tutti gli altri classificarli come "insuccessi".

Quando il numero di prove considerate N poi diviene molto grande ($N \to \infty$) e la probabilità di successo di un singolo evento p è piccola ($p \to 0$), purché il prodotto $m = Np$ si mantenga finito, allora la distribuzione binomiale tende ad assumere la forma della distribuzione di Poisson

$$P(n) = \frac{m^n}{n!} e^{-m}$$

La distribuzione di Poisson è cruciale nello studio statistico di eventi rari (per cui è piccola la probabilità di accadimento). Un modo alternativo per introdurre questo tipo di distribuzione è ricavare direttamente la legge di probabilitá. Studiamo ad esempio il caso del processo di decadimento di nuclei radioattivi. Stiamo osservando l'apparire di prodotti di decadimento di sostanze radioattive in intervalli di tempo successivi di ampiezza Δt. Ci chiediamo quale sia la legge che descrive la probabilita' di produzione di k eventi nel tempo di ossrvazione totale pari a t.
Assumiamo che
a) gli eventi di decadimento siano tra loro indipendenti
b) la probabilitá di avere un evento nell'intervallo di tempo Δt sia proporzionale a Δt stesso
c) la probailitá di avere piú di un evento in Δt sia un infinitesimo di ordine superiore rispetto a Δt.
Consideriamo due intervalli di tempo successivi $(0, t)$ e $(t, t + \Delta t)$ e scriviamo la probabilita' di non osservare eventi in ambedue gli intervalli. Per la condizione a) avremo che

$$P_o(0, t + \Delta t) = P_o(0, t) P_0(t, t + \Delta t) = P_o(t) P_0(\Delta t)$$

Sotto le ipotesi b) e c) potremo allora scrivere la probabilita‘ $P_0(\Delta t)$ in termini del suo complemento espresso dalla probabilita‘ $P_1(\Delta t) = \mu\Delta t$ di poter osservare un evento, $P_0(\Delta t) = 1 - \mu\Delta t$. Per cui si ha

$$P_o(t + \Delta t) = P_o(t)(1 - \mu\Delta t)$$

avremo

$$\frac{P_o(t + \Delta t) - P_o(t)}{\Delta t} = -\mu P_o(t)$$

Nel limite per $\Delta t \to 0$ otterremo l'equazione differenziale

$$\frac{dP_o(t)}{dt} = -\mu P_o(t)$$

che ammette come soluzione la funzione

$$P_o(t) = Ae^{-\mu t}$$

dove si vede subito che la costante d'integrazione A é pari ad 1 poiché é ovvio come in un tempo d'osservazione nullo non si possa osservare alcun evento di decadimento ($P_o(0) = 1$).
Iterando il precedente ragionamento ricaviamo allora la funzione $P_1(t)$. Nell'intervallo $(0, t + \Delta t)$ la probabilitá di avere un evento sará la somma della probabilitá di osservare un evento in Δt non avendolo osservato in $(0, t)$ e di osservare un evento in $(0, t)$ non avendolo osservato in Δt.

$$P_1(t + \Delta t) = P_0(\Delta t)P_1(t) + P_o(t)P_1(\Delta t)$$

$$P_1(t + \Delta t) = (1 - \mu\Delta t)P_1(t) + e^{-\mu t}P_1(\Delta t)$$

da cui, poiche' nel limite $\Delta t \to 0$ il rapporto $P_1(\Delta t)/\delta t)$ tende per l'ipotesi b) alla costante di proporzionalita' μ, avremo:

$$\frac{dP_1(t)}{dt} = -\mu P_1(t) + \mu e^{-\mu t}$$

Questa nuova equazione differenziale ammette come soluzione la funzione

$$P_1(t) = \mu t e^{-\mu t}$$

Iterando il procedimento per k eventi avremo

$$P_k(t) = \frac{(\mu t)^k}{k!} e^{-\mu t}$$

Si puo' dimostrare che per la distribuzione di Poisson ha valore aspettato pari a

$$E[k] = \sum_{k=o}^{\infty} k \frac{m^k}{k!} e^{-m} = m$$

e varianza

$$E[(k-m)^2] = m$$

DISTRIBUZIONI CONTINUE DI PROBABILITÀ: LA DISTRIBUZIONE DI GAUSS.

La distribuzione di Gauss fu introdotta da De Moivre alla fine del secolo XVIII^o. Gauss nel 1809 e Laplace nel 1812 la riproposero per descrivere la distribuzione degli errori in alcune misure astronomiche. La funzione di Gauss

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

è la densitá di probabilitá della variabile aleatoria continua **X** definita nello spazio degli eventi $(-\infty, +\infty)$.
Come abbiamo detto in precedenza, il differenziale $f(x)dx$ descrive la probabilitá che l'evento abbia associato un risultato compreso nell'intervallo $(x, x + dx)$. La funzione $f(x)$ è deducibile derivendo la sua primitiva $F(x)$, detta funzione di ripartizione. Tramite essa é possibile esprimere la probabilitá che l'evento abbiamo come risultato un valore compreso nell'intervallo finito (a, b).

$$F(b) - F(a) = \int_a^b f(x)dx$$

Coprendo tutto lo spazio dei possibili risultati, per definizione di probabilitá dobbiamo ottenere

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Applicando la definizione propria del caso delle distribuzioni continue di valor medio e varianza avremo per

la distribuzione di Gauss:

$$E[x] = \int_{-\infty}^{+\infty} x f(x) dx = \mu$$

$$E[(x-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx = \sigma^2$$

La funzione di Gauss ha la tipica forma di una campana; é simmetrica attorno al suo massimo che cade in $x = \mu$. La larghezza della campana dipende dal parametro σ, che descrive quantitativamente l'entitá della dispersione dei risultati attorno al valore aspettato μ. Integrando la funzione di Gauss nell' intervallo $(\mu - \sigma, \mu + \sigma)$ otteniamo la probabilitá che il risultato cada in questo stesso intervallo. Tale probabilitá risulta pari al 68,3 %. Per $(\mu - 2\sigma, \mu + 2\sigma)$ si ha il 95,6 % di probabilitá ed a $(\mu - 3\sigma, \mu + 3\sigma)$ il 99,7 %.