1 Statistica

Nelle lezioni precedenti abbiamo avuto qualche assaggio di statistica parlando di Graunt e di Arbuthnot.

Ma anche il metodo di Monte Carlo contiene qualche idea con gli esempi di Steinhaus e l'ago di Buffon.

In senso lato, stiamo usando la statistica tutte le volte che da un piccolo campione pensiamo di poter arrivare a delle conclusioni sull'insieme che ci troviamo a studiare.

Abbiamo una confezione di caramelle miste? Ne facciamo scivolare una decina sul tavolo e scopriamo che, tra queste, le caramelle alla menta sono il doppio di quelle alla liquirizia.

Possiamo congetturare allora che le caramelle sono solamente di due tipi e che prevalgono quelle alla menta.

Oppure andiamo in vacanza su un'isola caraibica, dall'aeroporto la sera ci facciamo portare direttamente all'albergo e la mattina dopo usciamo e vediamo che la maggior parte delle persone in strada sono bianche. Potremmo concluderne che gli abitanti del luogo sono per la maggior parte dei bianchi. Un'occhiata al sito internet o al libricino che ci hanno dato nell'agenzia turistica sarebbe più affidabile, ma magari un'informazione che ci interessa solo marginalmente e che facciamo così, senza pensarci troppo.

Una riflessione più attenta ci potrebbe magari far pensare che l'albergo a cinque stelle che abbiamo scelto sia nella parte più ricca della cittadina e che il "campione" che si è presentato alla nostra vista non è *rappresentativo*.

Elencherò alcuni degli errori più comuni che capita di riscontrare nelle indagini statistiche.

Il primo è quello del campione non rappresentativo. quando si fa un'indagine statistica su qualunque fenomeno (preferenze politiche, indagine di mercato su un certo prodotto, diffusione di una certa patologia, reddito) chi fa l'indagine deve prendere un campione che sia rappresentativo di tutte le categorie: per et, per genere, per residenza, per classe sociale, per rapporto tra grande città, città di provincia e campagna etc.

Un occhiata alla gente che passa davanti al nostro albergo può essere un esempio di campione non significativo: solo una visita più prolungata alla cittadina dove passiamo le vacanze potrà darci un'idea più precisa sulla presenza e consistenza della popolazione di colore.

Anche le caramelle sul piattino potrebbero non essere un campione non rappresentativo: nel negozio hanno forse messo nel sacchetto prima quelle alla liquirizia e poi quelle alla menta (che sono rimaste di sopra), oppure le caramelle sono incartate diversamente e quelle alla menta scivolano più facilmente dal sacchetto.

Spesso viene ricordato un famoso esempio storico, quello delle elezioni presidenziali americane del 1936. Una popolare rivista americana, The Literary Digest, inviò dieci milioni di lettere ai suoi potenziali lettori e ricevette 2.270.000 risposte, dalle quali emerse che il vincitore sarebbe stato Alf London, con uno scarto anche abbastanza netto. Contemporaneamente il giovane George Gallup, fondatore dell'omonima impresa, oggi famosa in tutto il mondo, scegliendo accuratamente un campione di 50.000 potenziali votanti, riuscì a prevedere la vittoria di Franklin Delano Roosevelt.

L'errore clamoroso del Literary Digest era dovuto a due concause: il primo era quello di aver scelto i dieci milioni di potenziali elettori unendo le liste degli abbonati, dei proprietari di automobile e di chi aveva il telefono.

L'automobile era, nel '36, segno di agiatezza e in parte anche l'abbonamento al telefono non era da tutti. Ma anche la rivista era diffusa soprattutto tra le classi benestanti che si potevano permettere l'abbonamento ai tempi della grande crisi economica.

Il clamore suscitato dall'insuccesso e dal confronto con il risultato ottenuto da Gallup portò, dopo un po', alla chiusura della rivista.

Ma il fallimento è stato così clamoroso che l'indagine è stata studiata anche successivamente negli anni '70 e '80. Da questi studi è emerso un altro errore, e cioè che le 2.270.000 risposte sono arrivate a persone che erano molto favorevoli a London, o molto contrariate da Roosevelt, mentre non avevano risposto piǔ di tre su quattro, ma che avevano probabilmente votato alla fin fine per il più tranquillizzante Roosevelt, presidente uscente.

L'analisi dei dati basati su un *campione autoselezionato*, cioè del campione si forma con la risposta, su base volontaria, di parte degli degli intervistati.

Un esempio (meno eclatante di quello della corsa presidenziale del '36) mi è stato segnalato da un neolaureato. Si sa che il consorzio interuniversitario AlmaLaurea raccoglie, tra tante informazioni, anche dati sull'impiego dei laureati: il periodo di attesa, la corrispondenza alla formazione ricevuta, lo stipendio. Basandosi sulla reazione dei suoi coetanei il mio ex studente ha congetturato che le risposte dianno una fotografia troppo ottimistica della

situazione a causa della riluttanza, da parte di chi è ancora in cerca di lavoro, o ha un lavoro sottopagato, di rispondere al questionario. O magari di rispondere abbellendo un po' la sua situazione.

Un altro errore frequente è quello del gruppo troppo piccolo. Siccome certe ricerche sono costose (efficacia di un certo medicinale, ad esempio) ci si accontenta di un piccolo numero di pazienti e il dato che se ne ricava non è significativo.

Un altro aspetto, piuttosto controverso, è il cosiddetto *effetto placebo*. Alcune condizioni patologiche sembrano sensibili alla somministrazione di sostanze inerti che però fanno credere ai pazienti di essere sottoposti ad una terapia. L'effetto dipende dalla condizione patologica. recenti ricerche hanno messo in dubbio una presenza generalizzata del fenomeno, confermandolo per il trattamento del dolore, il Parkinson, l'insonnia e la depressione. Le ricerche sono ancora in corso.

Un modo corretto di fare delle indagini statistiche è quello di fare un confronto tra due gruppi omogenei, confrontando gli effetti del trattamento con il medicinale già in uso e quelli ottenuti con la nuova terapia in esame (o anche l'effetto placebo).

Due ricercatori tedeschi (Hans-Hermann Dubben e Hans-Peter Beck-Bornholdt) hanno pubblicato nel 1997 un divertente libro intitolato *Der Hund der Eier legt (Il cane che fa le uova)* in cui presentavano una serie di esempi di uso errato della statistica anche nelle pubblicazioni scientifiche. Detto tra parentesi: meriterebbe una traduzione in italiano. Una parodia divertente è un grafico che mostra la correlazione significativa di due fenomeni tra il 1965 e il 1980 nella Germania Federale tra la natalità e la nidificazione delle cicogne: entrambi i fenomeni erano in discesa. Con questo volevano mettere in evidenza come alcune ricerche scientifiche confondono correlazione con causalità.

I dati sulla diffusione dell'attuale pandemia sono difficilmente leggibili. Veniamo sommersi di cifre prive di senso, perché il numero dei positivi, regione per regione, dipende dal numero dei tamponi effettuati che comunque, anche nelle regioni più solerti (diciamo così) sono insufficienti e autoselezionati. O selezionati con criteri che nulla hanno a che vedere con la sanità pubblica: i giocatori di serie A avranno fatto innumerevoli tamponi, mentre altre persone non riuscivano a farlo pur in presenza di difficoltà respiratorie.

I numeri relativi al Covid-19 siano privi di senso e che ci capiremo qualcosa

solo tra un anno o forse più. Questa riserva è suggerita da certe cifre diffuse dall'ISTAT che indicano che in certi comuni lombardi il numero dei decessi è aumentato più di quanto indichino le cifre sulla pandemia. I dati sono del tipo: nel comune X sono morte nel primo trimestre 2019 100 persone, nel 2020 sono aumentare a 400. Ufficialmente, per Covid-19, sono morte 200 persone. Ma allora i morti per altre cause sono 200, rispetto alle 100 dell'anno precedente? Non può trattarsi di una fluttuazione statistica. A bocce ferme si capirà meglio che cosa sia successo veramente.

Le statistiche dovranno essere analizzate anche negli anni a venire, perché si fa presto a dire che uno è "guarito". Sembra che molti sopravvivono alla terribile esperienza con danni permanenti ai polmoni, ma non solo, con speranza di vita probabilmente accorciata. Sarà importante analizzare anche nei prossimi anni le tavole di mortalità per fascia di età e per territorio. Tavole che la pandemia ha sconvolto e che andranno riviste.

Sulla pandemia infuria sul web una campagna destabilizzatrice utilizza i fatti a seconda delle convenienze.

Cito solamente un video che ho visto di recente e che ci riguarda più da vicino (come materia).

Ho sentito recentemente una persona che veniva qualificata come "medico di Udine" e che sosteneva che questa influenza è come le altre e che il governo (per motivi che non spiega) ha volutamente drammatizzato la situazione imponendo il lockdown e convogliando negli ospedali una massa di persone che andavano curate invece a casa (come?). A dargli ragione, diceva il "medico", è il fatto che quelli in ospedale sono morti !

Se la cosa non fosse tragica (e politicamente destabilizzante) sarebbe da ridere con Mark Twain, il quale disse "Il letto è il posto più pericoloso del mondo: vi muore l'80% della gente."

Riporto un altro esempio "sanitario" di cui sono stato testimone. Nel periodo che trascorsi presso l'università di Erlangen vidi un giorno i colleghi che stavano leggendo, con stupore ed eccitazione una notizia riportata sul giornale locale: nel comune di Erlangen la mortalità era superiore a quella di altre città tedesche. Cosa stanissima, perché Erlangen si vantava di essere una delle città più "verdi" della Germania: moltissime piste ciclabili, molti parchi nel centro della città, l'economia basata su centri di ricerca privati, una grande università e un grande ospedale. Un collega statistico si prese la briga di studiare la fonte della notizia. Scoprì così che il dato era dovuto all'aver iscritto, impropriamente, tra i deceduti a Erlangen coloro che erano

spirati nell'ospedale. L'ospedale era noto in tutta la Germania per il suo centro di ricerca e per la sperimentazione di nuove terapie. Vi arrivavano molti pazienti da fuori, anche dall'estero, e questi erano spesso in situazione critica. Dopo aver accertato quanti, tra i deceduti, erano residenti a Erlangen, ottenne il dato tutti si aspettavano: Erlangen era una città con un'altissima qualità di vita.

Le ricerche su dati sensibili sono particolarmente difficili. Ci sono delle domande alle quali l'intervistato in genere non risponde con sincerità: preferenze sessuali, evasione delle tasse, fedeltà coniugale, precedenti penali, etc). In questo caso le domande dirette rendono le risposte inutilizzabili. Per ciascun dato sensibile ci sono studi sulle domande indirette che si possono fare per ottenere informazioni ragionevoli sul fenomeno studiato.

Le risposte possono essere anche influenzate da motivi abbastanza futili. Se si vuol sapere qual è il rapporto tra i lettori dell'Espresso e di Chi, una domanda diretta rischia di sopravvalutare il numero dei lettori della prima rivista rispetto alla seconda. Sono state fatte delle indagini indirette, mandando nelle case degli intervistatori che si presentavano con la scusa che stavano raccogliendo carta da riciclare.

La statistica si presta anche a imbrogli e manipolazioni. Un famoso statista, Winston Churchill, disse: "mi fido solamente delle statistiche che ho personalmente manipolato".

C'è un'estesa letteratura su come si possa imbrogliare con la statistica. Nella mia biblioteca personale (oltre al testo tedesco citato prima) mi ritrovo "How to lie with statistics" di Darrell Huff oltre alla coppia di best seller "Damned lies and statistics" e "More damned lies and statistics" di Joel Best.

Le tecniche per imbrogliare con i numeri sono numerose. Alle volte trasparenti, altre volte non facili da individuare.

Un esempio, molto ingenuo, mi è rimasto in mente. Da anni varie istituzioni pubbliche crotonesi richiedono che venga potenziato e incluso nelle rotte nazionali l'aeroporto di Crotone, dove atterrano dei voli charter e forse anche altri voli locali. A causa della situazione sanitaria in cui ci troviamo, non sono riuscito a trovare dati recenti sull'operatività dell'aeroporto.

Ma tempo fa la stampa locale aveva riportato un dato a supporto delle richieste crotonesi: è risultato che gli aerei che decollavano da Crotone erano più pieni di quelli che decollavano da Lamezia!

Dato molto probabilmente vero, ma privo di significato. Quello che conta

è il potenziale mercato sul quale il nuovo aeroporto potrebbe contare tenendo conto delle distanze, della potenzialità dell'economia locale e della concorrenza degli aeroporti esistenti.

Un altro esempio che sono andato a cercare in rete, e ho naturalmente trovato, è la pubblicità ingannevole per alcuni prodotti di bellezza.

I vari effetti promessi dovrebbero essere supportati, per legge, da prove scientifiche, perciò alcune pubblicità riportano i risultati di "test autovalutativi" che provengono da qualche decina di persone. E già questo basta per squalificare la "ricerca". Inoltre l'opinione del campione potrebbe essere influenzata dal invio gratuito del prodotto o dalla manipolazione dei risultati: "riduzione delle rughe del 90%" potrebbe voler dire che 9 donne su 10 hanno apprezzato il prodotto.

Concludo questa parte introduttiva con un altro piccolo esempio che avevo notato sulla stampa nazionale una decina di anni fa. Un articolo trattava del problema del randagismo, presente in tutte le regioni italiane. Già trent'anni fa branchi di cani randagi si aggiravano nei boschi che circondano Trieste e creavano problemi a escursionisti e contadini.

Ebbene, questo articolo riportava una tabella con i dati relativi alle singole regioni. Quale era la regione con più cani randagi? Non lo credereste mai: era l'Emilia-Romagna.

Eppure il randagismo che salta gli occhi l'ho conosciuto nella mia vita durante i miei primi viaggi nel sud: mi avevano colpito i cani "stanziali" nel campus dell'università di Napoli a Monte Sant'Angelo. Quasi vent'anni fa mi sono trasferito all'Università della Calabria e ho visto quello che tutti voi conoscete.

Ma naturalmente il fenomeno del randagismo, come tutti gli altri problemi di questo mondo, non esiste se autorità competente se ne disinteressa e se non vengono fatti censimenti seri.

Concludiamo questo paragrafo introduttivo con un altro detto di Mark Twain: La gente di solito usa le statistiche come un ubriaco i lampioni: più per sostegno che per illuminazione.

Stimatori per la media e la varianza

Nella statistica ci troviamo il più delle volte nella seguente situazione: vogliamo studiare un fenomeno aleatorio, modellato da una v.a. X definita su uno spazio di probabilità $(\Omega, \mathcal{A}, P^{\vartheta})$.

La probabilità P^ϑ dipende da un parametro (reale o multidimensionale $\vartheta \in \Theta$) sul quale si cercano di raccogliere delle informazioni. La dipendenza dal parametro ϑ si rileva anche dalla densità di X che scriveremo $f(t;\vartheta)$.

Ripetiamo, in condizioni di indipendenza, l'esperimento. Questo equivale a considerare un'n-upla di copie *indipendenti* della X, $X_1, X_2, \ldots, X_n$ e dei valori osservati $x_1, x_2, \ldots, x_n$.

Potremmo ad esempio lanciare un dado 300 volte ed ottenere per i sei risultati possibili le seguenti frequenze:

$$53,\ \ 42,\ \ 49,\ \ 57,\ \ 62,\ \ 48\,.$$

Da questi risultati vorremmo farci un'idea sulla distribuzione di probabilità $(p_1, p_2, p_3, p_4, p_5, p_6)$ relativa a quel dado. In particolare ci interessa rispondere alla domanda se possiamo considerare che quel dado è regolare.

I dati raccolti si utilizzano per ottenere qualche informazione sulla X. Abbiamo visto, ad esempio, che il teorema di Glivenko-Cantelli ci permette di approssimare la funzione di ripartizione della X.

In questo paragrafo vedremo come si possono ottenere delle stime per la speranza matematica e la varianza della X. Più in generale, se la v.a dipende da un certo numero di parametri $\vartheta \in \Theta \subset \mathbb{R}^m$ (ad esempio una v.a. avente legge $\Gamma(\alpha, \lambda)$ dipende da due parametri), vogliamo stimare ϑ o, in certi casi, una funzione $\psi(\vartheta)$ di ϑ. Vedremo un esempio di questo tipo quando parleremo della v.a. esponenziale X.

In senso lato si chiama *stimatore* di un parametro ϑ una qualunque funzione del campione, che possiamo indicare con $T = t(X_1, X_2, \ldots, X_n)$. Si tratta ovviamente di una variabile aleatoria.

Se non facciamo ulteriori precisazioni sulla funzione T questa definizione è decisamente troppo ampia.

Diremo che uno stimatore è *non distorto* (o anche corretto) di $\psi(\vartheta)$ se per ogni $\vartheta \in \Theta$

$$E^\vartheta(T) = \psi(\vartheta)\,.$$

Naturalmente T, come ogni variabile aleatoria, può assumere diversi valori, ma il valore medio atteso è uguale al parametro (o a una sua funzione) che desideriamo stimare.

Media empirica

Se $X_1, X_2, \ldots, X_n$ è un campione relativo alla v.a. X, allora si dice *media empirica* la v.a. $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

È facile vedere che si tratta di uno stimatore non distorto della speranza matematica di X.

Infatti,

$$E^{\vartheta}(\bar{X}) = E^{\vartheta}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E^{\vartheta}(X_i) = E^{\vartheta}(X)\,.$$

Gli stimatori dipendono naturalmente dalla numerosità del campione. In generale si omette di indicare la dipendenza da n, per non appesantire la notazione. Ma, se può essere utile, si può scrivere T_n invece di T.

Uno stimatore per $\psi(\vartheta)$ si dice *coerente*, se $\lim_{n\to\infty} E^{\vartheta}(T_n) = \psi(\vartheta)$

Nel paragrafo successivo vedremo un esempio di stimatore coerente.

Stimatori per la varianza

Sia, al solito, $X_1, X_2, \ldots, X_n$ è un campione relativo alla v.a. X e supponiamo di conoscere μ, la media attesa di X.

È una condizione che si verifica raramente, ma in questo momento è comodo per noi fare questa ipotesi.

In questo caso si vede facilmente che $\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ è uno stimatore non distorto della varianza di X.

Infatti

$$E^{\vartheta}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n} E^{\vartheta}((X_i-\mu)^2) =$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Var\ X} = \mathrm{Var\ X}\,.$$

Di solito però non si conosce μ. Allora si utilizza un metodo empirico che spesso si rivela utile.

Non conoscendo μ, ne prendiamo un surrogato e cioè $\bar{X}$ e consideriamo lo stimatore

$$\bar{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\,.$$

Si ha che

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2\bar{X}X_i + \bar{X}^2) =$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\bar{X}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\bar{X}^2 + \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 \,.$$

Calcoliamo la speranza matematica di quest'ultima espressione (qui $\vartheta = (\mu, \sigma^2)$)

$$E^\vartheta\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right) = \frac{1}{n}\sum_{i=1}^{n} E^\vartheta(X_i^2) - \frac{1}{n^2}\left(\sum_{i=1}^{n} E^\vartheta(X_i^2) + \sum_{j\neq k} E^\vartheta(X_j X_k)\right) =$$

$$\frac{1}{n}\sum_{i=1}^{n} E^\vartheta(X_i^2) - \frac{1}{n^2}\left(nE^\vartheta(X^2) + \sum_{j\neq k} E^\vartheta(X_j)E^\vartheta(X_k)\right) =$$

$$E^\vartheta(X^2) - \frac{1}{n}E^\vartheta(X^2) + \frac{n^2-n}{n^2}E^\vartheta(X)^2 =$$

$$\frac{n-1}{n}(E^\vartheta(X^2) - E^\vartheta(X)^2) = \frac{n-1}{n}\mathrm{Var}^\vartheta(X)\,.$$

Ne segue che $\bar{S}^2$ non è uno stimatore non distorto. È però un esempio di stimatore coerente.

Comunque, a conti fatti, non è difficile, servendoci di $\bar{S}^2$, costruire uno stimatore non distorto. Basta considerare

$$S^2 = \frac{n}{n-1}\bar{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\,.$$

Stimatori di massima verosimiglianza

Sia $X_1, X_2, \ldots, X_n$ un campione relativo alla v.a. X avente densità $f(t;\vartheta)$.

La funzione $L(t_1, t_2, \ldots, t_n; \vartheta) = \prod_{k=1}^{n} f(t_k;\vartheta)$ si chiama, in questo contesto, funzione di verosimiglianza.

Lo *stimatore di massima verosimiglianza* è quel valore di ϑ (se esiste) che massimizza la probabilità dei valori osservati.

Lo stimatore si calcola massimizzando rispetto a ϑ la densità congiunta nella quale i valori t_i sono noti, essendo i risultati del campionamento.

Per semplificare i calcoli, conviene massimizzare piuttosto la funzione

$$l(t_1, t_2, \ldots, t_n; \vartheta) = \log\prod_{k=1}^{n} f(t_k;\vartheta) = \sum_{k=1}^{n}\log f(t_k;\vartheta)\,.$$

Le due funzioni L ed l hanno gli stesi massimi, essendo il logaritmo strettamente crescente.

Esempi

1) V.a. di Poisson

Sia $X_1, X_2, \ldots, X_n$ un campione relativo ad una variabile di Poisson della quale non conosciamo il parametro λ.

Sappiamo che $P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$. La funzione di verosimiglianza è

$$\prod_{i=1}^{n} \frac{\lambda^{k_i}}{k_i!}e^{-\lambda}\,.$$

Il suo logaritmo è

$$\sum_{i=1}^{n}(k_i \log\lambda - \lambda - \log k_i!)$$

Usiamo il calcolo differenziale e calcoliamo la derivata (rispetto a λ). Otteniamo

$$l'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} k_i - n\,.$$

Si vede subito che $l'(\lambda) = 0$ se e solo se

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} k_i\,,$$

abbiamo cioè ritrovato la media empirica. Poiché la funzione è concava, si tratta effettivamente di un massimo.

2) Variabile aleatoria normale $N(\mu, \sigma^2)$.

In questo caso il parametro da stimare è bidimensionale, $\vartheta = (\mu, \sigma^2)$.

La funzione di verosimiglianza è

$$L(t_1, t_2, \ldots, t_n; \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(t_i - \mu)^2)$$

e il suo logaritmo è

$$l(t_1, t_2, \ldots, t_n; \mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(t_i - \mu)^2\,.$$

Calcoliamo le derivate parziali.

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(t_i - \mu)\,.$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(t_i - \mu)^2\,.$$

Annullando le derivate si ottiene

$$\sum_{i=1}^{n}(t_i - \mu) = 0\,, \;\text{ cioé }\; \mu = \frac{1}{n}\sum_{i=1}^{n} t_i \quad \text{e}$$

$$-n\sigma^2 + \sum_{i=1}^{n}(t_i - \mu)^2 = 0\ , \;\text{ cioé }\; \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(t_i - \mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(t_i - \bar{X})^2\,.$$

Si verifica facilmente che il punto critico trovato è effettivamente un punto di massimo.

In conclusione, μ viene stimato con la media empirica, mentre σ^2 viene stimato con lo stimatore distorto, ma coerente, che abbiamo trovato nel paragrafo precedente.

3) Vediamo ora un esempio per il quale il metodo non dà risultati soddisfacenti.

Sia X la variabile aleatoria uniformemente distribuita sull'intervallo $I_\delta = [-\delta, \delta]$.

La densità di X è la funzione $f(t;\delta)$ che vale $\frac{1}{2\delta}$ su I_δ e 0 altrimenti.

La funzione di verosimiglianza è (e qui posiamo tralasciare il logaritmo)

$$L(t_1, t_2, \ldots, t_n; \delta) = \prod_{i=1}^{n} \mathbb{1}_{I_\delta}(t_i)\,.$$

Nel disegno sottostante vediamo 20 punti osservati.

—.–.——.—.–..—.–.–.—-0—.–.—.—.——..—-.–..—-.—-

Al crescere di δ la funzione di verosimiglianza rimane nulla finché qualche fattore rimane nullo. Quando $\tilde{\delta} = \max_i |t_i|$, L ha una discontinuità e da quel punto in poi si mantiene costantemente uguale a 1.

Possiamo prednere $\tilde{\delta}$ come stimatore, il quale però tende a sottostimare il parametro cercato.

Si può dimostrare che pur non essendo uno stimatore non distorto, esso è coerente.

Il metodo dei momenti

Sia $X_1, X_2, \ldots, X_n$ un campione relativo alla v.a. X che dipende da k parametri (cioé $\Theta \subset \mathbb{R}^k$) e ammetta momenti almeno fino all'ordine k.

Indichiamo con $m_i(\vartheta) = E^{\vartheta}(X^i)$ il momento i-esimo di X.

Indichiamo invece con $\tilde{m}_i = \frac{1}{n}\sum_{j=1}^{n} X_j^i$ il *momento empirico* i-esimo.

Per applicare il metodo dei momenti, bisogna risolvere il sistema

$$
\begin{aligned}
&m_1(\vartheta_1, \vartheta_2, \ldots, \vartheta_k) = \tilde{m}_1 \\
&m_2(\vartheta_1, \vartheta_2, \ldots, \vartheta_k) = \tilde{m}_2 \\
&\ldots \\
&m_k(\vartheta_1, \vartheta_2, \ldots, \vartheta_k) = \tilde{m}_k
\end{aligned}
$$

Se $k = 1$ le cose sono abbastanza semplici.

Vediamo qualche esempio.

1) Supponiamo che X sia la v.a. esponenziale di parametro λ.

Sappiamo che $m_1(\lambda) = \frac{1}{\lambda}$, mentre $\tilde{m} = \frac{1}{n}\sum_{j=1}^{n} X_j = \bar{X}$ e quindi l'equazione $m_1 = \tilde{m}_1$ implica che

$$\frac{1}{\lambda} = \bar{X}$$

e quindi uno stimatore per λ è

$$\tilde{\lambda} = \frac{1}{\bar{X}}\,.$$

2) Sia $X \sim B(n, p)$. Sappiamo che $m_1 = E(X) = np$ e $m_2 = E(X^2) = npq + n^2p^2$.

Dalla seconda equazione deduciamo che $m_2 = m_1(1 - p + m_1)$, da cui si deduce che

$$m_2 - m_1^2 = m_1(1 - p)$$

da cui

$$\hat{p} = 1 - \frac{m_2 - m_1^2}{m_1} \text{ e } \hat{n} = m_1 - m_2 - m_2^2 \,.$$

È molto probabile che $\hat{n}$ non sia un intero. In questo caso prendiamo come stimatore di n l'intero più vicino a $m_1 - m_2 - m_2^2$.

3) Sia X uniformemente distribuita su $[0, a]$. Vogliamo trovare uno stimatore per a.

Sappiamo che $E(X) = a/2$ e quindi $a = 2m_1$.

Lo stimatore dei momenti è $\hat{a} = 2\bar{X}$. Supponiamo di aver rilevato i seguenti dati: $1.1, 2.3, 2.9$ e 9.7. Allora $\bar{X} = 4$ e $\hat{a} = 8$.

Si noti che lo stimatore ci dà un risultato sicuramente sballato, poiché tra i valori osservati c'è anche 9.7, che è maggiore, e non di poco, della stima $\hat{a}$ ottenuta. Naturalmente la colpa non è dello stimatore scelto, ma della ridotta dimensione del campione.

4) Studiamo ora la v.a. uniformemente distribuita sull'intervallo $[a, b]$.

Sappiamo che $E(X) = \frac{a+b}{2}$ e che $E(X^2) = \frac{1}{b-a}\int_a^b t^2\,dt = \frac{1}{3}(b^2 + ab + a^3)$.

Il sistema da studiare è allora

$$\tilde{\mu}_1 = \frac{a+b}{2}$$

$$\tilde{\mu}_2 = \frac{1}{3}(b^2 + ab + a^2)\,.$$

Usiamo un po' di algebra. Dalla prima equazione ricaviamo $a = 2\tilde{\mu}_1 - b$. Sostituendo questa espressione nella seconda equazione, otteniamo

$$(2\tilde{\mu}_1 - b)^2 + b^2 + (2\tilde{\mu}_1 - b)b = 3\tilde{\mu}_2\,,$$

ovvero

$$b^2 - 4\tilde{\mu}_1 b + 4\tilde{\mu}_1^2 + b^2 + 2b\tilde{\mu}_1 - b^2 = 3\tilde{\mu}_2\,,$$

semplificando (opportunamente)

$$3\tilde{\mu}_1^2 + \tilde{\mu}_1^2 - 2b\tilde{\mu}_1 + b^2 = 3\tilde{\mu}_2$$

che possiamo riscrivere come

$$(b - \tilde{\mu}_1)^2 = 3(\tilde{\mu}_2 - \tilde{\mu}_1^2)$$

e siccome $b > \tilde{\mu}_1$,

$$\hat{b} = \tilde{\mu}_1 + \sqrt{3(\tilde{\mu}_2 - \tilde{\mu}_1^2)} \quad \text{e}$$

$$\hat{a} = \tilde{\mu}_1 - \sqrt{3(\tilde{\mu}_2 - \tilde{\mu}_1^2)}\,.$$

Simulando con l'iphone dieci numeri casuali in $[0, 1]$, ho ottenuto le stime $\tilde{\mu}_1 = 0,396$ e $\tilde{\mu}_2 = 0,166$. Inserendo questi valori nelle equazioni precedenti troviamo le stime

$$\hat{a} = -0,124 \text{ e } \hat{b} = 0,92\,.$$

Intervalli di fiducia

Nei paragrafi precedenti abbiamo cercato delle stime puntuali per i parametri incogniti. In questa sezione studiamo il problema della stima da un'altra angolazione.

Se $T_1 = t_1(X_1, X_2, \ldots, X_n)$ e $T_2 = t_1(X_1, X_2, \ldots, X_n)$ sono due stimatori per $\psi(\vartheta)$, diciamo che l'intervallo $J_X = [T_1, T_2]$ è un intervallo di fiducia di livello $1 - \alpha$ per $\psi(\vartheta)$, se per ogni $\vartheta \in \Theta$ si ha che

$$P^{\vartheta}(\psi(\vartheta) \in J_X) \geq 1 - \alpha\,.$$

Si noti che gli estremi, e quindi l'intervallo, sono aleatori, mentre non è aleatorio $\psi(\vartheta)$.

Quando abbiamo parlato dell'approssimazione normale, abbiamo fatto delle considerazioni che ci portano ad un primo esempio.

Supponiamo di lanciare in dado (non necessariamente regolare) 2400 volte e di ottenere 360 volte il 6. Quindi è $\bar{X} = 0,15$. Indichiamo con p la probabilità che con quel dado si ottenga 6.

Vogliamo trovare un intervallo che contiene p con probabilità di $0,95$.

Useremo l'approssimazione normale.

$$P(|\bar{X} - p| > \eta) \sim 2\Phi\left(-\frac{\eta}{\sqrt{p(1-p)}}\sqrt{2400}\right)\,.$$

Possiamo maggiorare $\sqrt{p(1-p)}$ con $\frac{1}{2}$ e quindi da

$$P(|\bar{X} - p| > \eta) \sim 2\Phi(-2\eta\sqrt{2400}) = 2(1 - \Phi(-2\eta\sqrt{2400})) = 0,05\,.$$

deduciamo che

$$\Phi(2\eta\sqrt{2400})) = 0.975$$

Guardando alla tabella della normale otteniamo che

$$2\eta\sqrt{2400} = 1,96 \text{ da cui } \eta \sim \frac{1,96}{2 \cdot 49} = 0,2\,.$$

Ne deduciamo che un intervallo di fiducia del livello del 95% per p è dato (circa) da $[-0,05, 0,35]$.

Il risultato non è molto soddisfacente. L'estremo sinistro dell'intervallo è addirittura negativo! Vedremo come si può fare di meglio.

Stima intervallare per campioni gaussiani

Conviene introdurre ora il concetto di *quantile* di una v.a. X avente funzione di ripartizione F_X.

Fissiamo un numero $\alpha \in]0,1[$. Chiameremo *quantile di ordine* α della v.a. X il più grande numero reale q_α tale che

$$F_X(q_\alpha) = P(X \leq q_\alpha) \leq \alpha\,.$$

Se F_X è continua, allora esiste almeno un q_α che soddisfa all'equazione $F_X(q_\alpha) = \alpha$.

Possiamo farci un'idea sui quantili dando nuovamente un'occhiata alla densità della normale standard

https://www.mathsisfun.com/data/standard-normal-distribution-table.html

Se, come capita nei casi in cui utilizzeremo questo concetto, la funzione di ripartizione della X è strettamente crescente (tranne che, eventualmente, su due semirette $]-\infty, a]$ e/o $[b, +\infty[$ dove $F_X(t) \equiv 0$ e/o $F_X(t) \equiv 1$, rispettivamente), esiste ed è unico il numero reale q_α tale che

$$F_X(q_\alpha) = \alpha\,.$$

Se la X ha legge $N(0,1)$, indichiamo con Φ_α il suo quantile di ordine α.

Poiché la densità di X è pari, X e $-X$ hanno la stessa legge e risulta, come abbiamo visto, $\Phi(t) = 1 - \Phi(-t)$, da cui segue che

$$P(X \leq -\Phi_\alpha) = P(-X \leq -\Phi_\alpha) = P(X \geq \Phi_\alpha) = 1 - P(X \leq \Phi_\alpha) = 1 - \alpha$$

e quindi

$$-\Phi_\alpha = \Phi_{1-\alpha}\,.$$

Un'altra identità che ci sarà utile è la seguente:

$$P(|X| \leq \Phi_{1-\frac{\alpha}{2}}) = P(-\Phi_{1-\frac{\alpha}{2}} \leq X \leq \Phi_{1-\frac{\alpha}{2}}) =$$

$$P(X \leq \Phi_{1-\frac{\alpha}{2}}) - P(X \leq -\Phi_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha\,.$$

Consideriamo un campione $X_1, X_2, \ldots, X_n$ di una v.a. di legge $N(\mu, \sigma^2)$.

Supponiamo che si conosca il valore di σ^2. È un caso raro, ma utile per i suggerimenti che può dare per sviluppi più complessi.

Nelle considerazioni preliminari al TCL abbiamo osservato che la v.a. $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ ha legge $N(0,1)$.

Ne segue che

$$1-\alpha = P^{\vartheta}\left(\left|\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu)\right| \leq \Phi_{1-\frac{\alpha}{2}}\right) = P^{\vartheta}\left(-\Phi_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}}{\sigma}(\bar{X}-\mu) \leq \Phi_{1-\frac{\alpha}{2}}\right) =$$

$$P^{\vartheta}\left(\bar{X} - \frac{\sigma}{\sqrt{n}}\Phi_{1-\frac{\alpha}{2}} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}}\Phi_{1-\frac{\alpha}{2}}\right),$$

dunque

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}}\Phi_{1-\frac{\alpha}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}}\Phi_{1-\frac{\alpha}{2}}\right]$$

è un intervallo di fiducia di livello $1-\alpha$ per il parametro μ.

Di solito però non conosciamo la varianza. Nel migliore dei casi possiamo maggiorarla, come abbiamo fatto in un esempio precedente. Ma questo capita raramente.

L'esempio precedente ci suggerisce però un percorso che già abbiamo seguito quando abbiamo cercato uno stimatore per la varianza, non conoscendo il valore della speranza matematica μ. Abbiamo allora empiricamente sostituito μ con un suo stimatore, $\bar{X}$, salvo fare poi, alla fine dei conti, una verifica per vedere se lo stimatore così ottenuto era non distorto oppure no.

In questo caso procediamo in modo analogo. Non conoscendo σ^2, l'idea è quella di prendere al suo posto un suo stimatore non distorto, S^2, e di considerare la v.a.

$$Z = \frac{\sqrt{n}}{\sqrt{S^2}}(\bar{X}-\mu)$$

e di trovare un intervallo di fiducia per μ cercando di rifare il percorso fatto prima.

Si può dimostrare che Z ha legge di Student con $n-1$ "gradi di libertà". Questa legge si denoterà con $t(n-1)$. La sua densità è

$$\frac{\Gamma(\frac{n+1}{2})}{\sqrt{2\pi}\,\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} .$$

Si tratta di una funzione pari. Per $n=1$ si ritrova la v.a. di Cauchy.

In generale, come si vede, a parte una costante, la densità di $t(n)$ è

$$c_n\left[\left(1+\frac{t^2}{2n}\right)^{-n}\right]^{\frac{1}{2}}\left(1+\frac{t^2}{2n}\right)^{-\frac{1}{2}} .$$

Il secondo fattore tende, al tendere di n all'infinito, a $e^{-\frac{t^2}{2}}$, mentre il terzo tende ovviamente a 1. Ne consegue che

$$\frac{\Gamma(\frac{n+1}{2})}{\sqrt{2\pi}\,\Gamma(\frac{n}{2})} \quad \text{tende a} \quad \frac{1}{\sqrt{2\pi}}$$

e che, se $X_n \sim t(n)$, allora la successione tende, in legge, ad una $N(0,1)$.

Il grafico della densità di $t(n)$ ha un comportamento che ricorda quello della normale standard, avendo però un valore nello 0 un po' più piccolo ed avendo le "code" più spesse. Infatti, come si vede dall'espressione analitica, la densità di $t(n)$ ha un comportamento asintotico equivalente a $t^{-(n+1)}$.

I quantili della legge $t(n)$ si trovano in tabelle per certi versi simili a quelle della normale standard.

Per quanto osservato prima, le tabelle riportano solamente i valori dei quantili per n relativamente piccolo (fino a 40 o 60), per poi passare a "∞", intendendo con questo la $N(0,1)$.

Dai grafici delle densità delle $t(n)$ e della $N(0,1)$ abbiamo una conferma (ma non ce n'era bisogno) di questo fatto. Il sito di Wikipedia

https://en.wikipedia.org/wiki/Student%27s_t-distribution

illustra le proprietà della legge, riporta i grafici di alcune densità significative e si chiude con le tabelle dei quantili significativi.

Denoteremo con $t_\alpha(n)$ i quantili della $t(n)$. Poiché la densità è una funzione pari, valgono proprietà analoghe a quelle che abbiamo visto per i quantili della $N(0,1)$.

In particolare, se $Z \sim t(n)$,

$$P(Z \leq -t_{1-\alpha}(n)) = \alpha$$

$$P(|Z| \leq t_{1-\alpha/2}(n)) = 1 - \alpha\,.$$

Esempi

1) Supponiamo che su 20 campioni di birra venga misurata la gradazione alcolica. La media risulta essere di $5,3\%$, con uno scarto quadratico medio di $0,25$ gradi.

Vogliamo determinare per la gradazione alcolica di quella partita di birra un intervallo di confidenza di livello del 95%.

Per un α generico l'intervallo è

$$\left[5,3 - \frac{0,25}{\sqrt{20}} t_{1-\alpha/2}(19), 5,3 - \frac{0,25}{\sqrt{20}} t_{1-\alpha/2}(19)\right]\,.$$

Per rispondere alla nostra domanda basta trovare sulla tabella il valore di $t_{0.975}(19) = 2,093$. A conti fatti, l'intervallo cercato è

$$[5.183, 5.417]\,.$$

Se invece vogliamo un intervallo corrispondente al livello di fiducia del 99%, si ha che $t_{0.99}(19) = 2.861$ e con calcoli semplici si trova

$$[5.14, 5.46]\,.$$

L'intervallo è più grande, come potevamo prevedere, perché maggiore è la sicurezza richiesta, più ampio è l'intervallo.

Uno studente, prima dell'esame, si dice sicuro di prendere 30. Ma l'amico gli chiede: sei disposto a scommettere un caffé? Al che risponde: beh, ma almeno 27. L'amico lo incalza: scommetti una pizza? La sicurezza richiesta è sempre più elevata e quindi lo studente dice: scommettiamo una pizza che prendo almeno 25?

L'intervallo di fiducia non è unico, nel senso che alle volte interessa conoscere la probabilità (con un certo livello di fiducia) che la media che

ci interessa non sia superiore ad un certo valore o che non sia inferiore ad un certo valore.

2) Sappiamo che per legge il mercurio presente nel pesce in vendita non può superare i 5mg per un chilogrammo. Supponiamo di aver preso 10 campioni di pesce da un peschereccio e di aver misurato una media di 4.3 mg/kg e uno scarto quadratico medio di 0.9 mg/kg. Vogliamo determinate un intervallo di fiducia di livello del 99% del tipo $]-\infty, b]$

Per quanto visto, l'estremo destro dell'intervallo sarà $4.3 + \frac{0.9}{3.16} \cdot t_{0,99}(9)$.

Essendo $t_{0,99}(9) = 2.83$, l'intervallo cercato è

$$]-\infty, 5.1]\,,$$

quindi a quel livello di fiducia c'è la possibilità che parte di quel pesce non sia da mettere in vendita.

3) Perché un vino rosso sia commerciabile deve avere almeno 10 gradi.

In una cantina ci sono dodici botti contenenti lo stesso vino. Da ciascuna viene si preleva la quantità necessaria e se ne misura il grado alcolico.

Si ottiene la media di 10.5% con uno scarto quadratico medio di 0.3%. Vogliamo ottenere un intervallo di fiducia di livello 0.95 per la gradazione di quel vino del tipo $[a, +\infty[$.

Le considerazioni che abbiamo visto ci dicono che l'estremo sinistro dell'intervallo cercato sarà $10.5 - \frac{0.3}{3.46} \cdot t_{0,95}(11)$.

Poiché $t_{0,95}(11) \sim 1.796$, l'estremo sinistro risulta $10.5 - 0.16 = 10.34$ e quindi il vino può essere messo in vendita.

Confronto di stimatori

Si dice che uno stimatore T ha *varianza finita*, se $E^\vartheta(T^2) < +\infty$ per ogni $\vartheta \in \Theta$. Tutti gli stimatori che abbiamo visto e che vedremo godono di questa proprietà. Consideriamo uno stimatore T di varianza finita del parametro $\psi(\vartheta)$.

Si chiama *rischio quadratico medio* di T la funzione

$$R_T(\vartheta) = E^\vartheta((T-\vartheta)^2)\,.$$

Possiamo scrivere

$$R_T(\vartheta) = E^\vartheta\left[\big((T - E^\vartheta(T)) + (E^\vartheta(T) - \vartheta)\big)^2\right] =$$

$$= E^\vartheta((T - E^\vartheta(T))^2) + (E^\vartheta(T) - \vartheta)^2 + 2(E^\vartheta(T) - \vartheta)E^\vartheta(T - E^\vartheta(T)) =$$
$$= E^\vartheta((T - E^\vartheta(T))^2) + (E^\vartheta(T) - \vartheta)^2 \,,$$

tenuto conto del fatto che $E^\vartheta(T - E^\vartheta(T)) = 0$.

D'altra parte il primo addendo è proprio $\mathrm{Var}_\vartheta(T)$ e quindi $R_T(\vartheta)$ è la somma della varianza di T e del quadrato della differenza tra il valore medio atteso di T e del parametro ϑ da stimare.

Se lo stimatore è non distorto il secondo addendo si annulla e si ha che

$$R_T(\vartheta) = \mathrm{Var}_\vartheta(T) \,.$$

La varianza è una misura della dispersione dei valori intorno al suo valor medio atteso. Se in particolare T non è distorto, possiamo dire che più piccola è la varianza di T, più sono concentrati, in media, i suoi valori intorno al parametro ϑ.

È naturale preferire uno stimatore T' ad uno stimatore T'' se

$$R_{T'}(\vartheta) \leq R_{T''}(\vartheta)$$

per ogni $\vartheta \in \Theta$.

C'è da tener conto però del fatto che il rischio quadratico medio è una funzione di ϑ e che tra le funzioni non c'è un ordine totale, e che quindi ci possono essere degli stimatori tra loro non confrontabili, almeno non con questo criterio.

Esempio

Sia $X_1, X_2, \ldots, X_n$ un campione relativo ad una v.a. X. Sappiamo che la media empirica $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ è uno stimatore non distorto della speranza matematica di X.

Altri stimatori non distorti di $E^\vartheta(X)$ si possono ottenere (per esempio, ma ci sono anche altri modi) prendendo una qualunque n-upla di numeri reali positivi $a_1, a_2, \ldots, a_n$ tali che $\sum_{i=1}^n a_i = 1$ e considerando lo stimatore $\hat{X} = \sum_{i=1}^n a_i X_i$.

Faremo vedere che $\bar{X}$ è preferibile a $\hat{X}$.

Intanto dimostriamo che $\hat{X}$ è non distorto. Risulta infatti

$$E^\vartheta(\hat{X}) = E^\vartheta\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i E^\vartheta(X_i) = \sum_{i=1}^n a_i E^\vartheta(X) = E^\vartheta(X) \,.$$

Possiamo allora confrontare direttamente le rispettive varianze. Risulta

$$\mathrm{Var}_\vartheta\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}_\vartheta(X_i) = \frac{1}{n}\mathrm{Var}_\vartheta(X)\,,$$

mentre si ha

$$\mathrm{Var}_\vartheta\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \mathrm{Var}_\vartheta(X_i) = \mathrm{Var}_\vartheta(X)\sum_{i=1}^n a_i^2\,.$$

Si può ora dimostrare, usando ad esempio il metodo dei moltiplicatori di Lagrange, che la funzione delle n variabili $\varphi(a_1, a_2, \ldots, a_n) = \sum_{i=1}^n a_i^2$, con il vincolo $\sum_{i=1}^n a_i = 1$ assume il valore minimo quando $a_i = \frac{1}{n}$ per ogni $1 \le i \le n$.

Vogliamo determinare il minimo della funzione $\varphi(a_1, a_2, \ldots, a_n) = \sum_{i=1}^n a_i^2$ *soggetta al vincolo* $g(a_1, a_2, \ldots, a_n) = \sum_{i=1}^n a_i - 1 = 0$.

Il metodo dei moltiplicatori di Lagrange consiste nel trovare i punti critici della funzione

$$F(a_1, a_2, \ldots, a_n, \lambda) = \varphi(a_1, a_2, \ldots, a_n) - \lambda\, g(a_1, a_2, \ldots, a_n)\,.$$

Risulta

$$\frac{\partial F}{\partial a_i} = 2a_i - \lambda \quad e$$

$$\frac{\partial F}{\partial \lambda} = \sum_{i=1}^n a_i - 1\,.$$

Annullando le derivate rispetto ad a_i *deduciamo che per ogni* $1 \le i \le n$ $a_i = \frac{\lambda}{2}$ *e quindi, tenuto conto del vincolo, che l'unico punto critico è l'*n*-upla* $(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$.

Considerazioni geometriche ci portano a concludere che si tratta di un minimo e che quindi $\bar{X}$ *è preferibile a* $\hat{X}$*, qualunque sia la scelta degli* $a_i \ge 0$ *(con il vincolo che sappiamo).*

I calcoli fatti ci permettono anche di affermare che se aggiungiamo al campione un'ulteriore copia indipendente della v.a. su cui stiamo indagando, X_{n+1}, allora

$$\mathrm{Var}_\vartheta(\bar{X}_n) = \frac{1}{n}\mathrm{Var}_\vartheta(X) > \frac{1}{n+1}\mathrm{Var}_\vartheta(X) = \mathrm{Var}_\vartheta(\bar{X}_{n+1})\,,$$

e che quindi la varianza di $\bar{X}_n$ decresce al crescere di n.

Vedremo ora qualche esempio che introduce al tema successivo al quale ci interesseremo. Il primo esempio è storico ed è interessante di per sé.

Gregor Mendel

Gregor Mendel (1822-1884) era un monaco austriaco di lingua tedesca, laureato in matematica e biologia.

Studiò prima nella piccola, ma antica, università di Olomouc e poi a Vienna, dove ebbe tra i professori il noto fisico Doppler.

Come uno dei migliori studenti, fu nominato assistente all'istituto di fisica.

Finiti gli studi, entrò in un convento vicino a Brno (oggi Repubblica Ceca), dove trovò un ambiente molto favorevole agli studi.

Nei due ettari dell'orto del suo monastero fece per molti anni studi sulla pianta del pisello commestibile, facendo esperimenti su varie caratteristiche genetiche utilizzando anche strumenti di probabilità e di statistica.

In tanti anni di esperimenti osservò sette caratteristiche dei piselli che riguardavano la forma del seme: liscio o rugoso, il colore del seme: giallo o verde, e altri aspetti che riguardavano la forma e il colore del baccello, il colore dei fiori e la loro posizione, lungo il fusto o in cima, e infine la lunghezza dei fusti.

Una di queste caratteristiche era il colore dei fiori.

I piselli del suo orto erano di due tipi: con i fiori viola oppure bianchi.

Incrociando tra di loro piante aventi fiori bianchi con se stesse, otteneva solamente piante a fiori bianchi (chiamò omozigote queste piante). Niente di interessante su quel fronte.

Autoinseminando invece i fiori dei piselli viola, notò che queste erano in effetti di due tipi: alcune avevano come discendenti esclusivamente piante con fiori viola (e la cui discendenza consisteva esclusivamente di piante con fiori viola). Chiamò omozigote anche queste.

Altre avevano invece come discendenti alcune piante con i fiori viola ed altre con i fiori bianchi. Chiamò eterozigote queste piante.

Osservò anche che le piante eterozigote generavano piante con fiori viola e con fiori bianchi bianche in proporzione più o meno costante di 3:1.

Questa osservazione gli fece ritenere che le piante eterozigote avessero in sé "traccia" di entrambi i caratteri. L'allele corrispondente al carattere

dominante determinava l'aspetto esteriore del discendente (viola nel caso che stiamo esaminando), ma quello latente rimaneva nel partimonio genetico di parte della discendenza e poteva riemergere nelle generazioni successive.

Ogni individuo ha, per ogni caratteristica (chiamiamola $\mathcal{A}$), nel suo patrimonio genetico due alleli. Indicheremo con A l'allele che è portatore del carattere dominante e con a quello recessivo. Un individuo può avere due alleli dello stesso tipo, AA o aa (sono gli omozigoti) oppure di tipo diverso Aa (e sono gli eterozigoti).

Incrociandosi con il "partner" (nel caso dei fiori il polline è maschile, il pistillo è femminile) cede al "discendente" uno dei due alleli.

E qui Mendel utilizzò i concetti di probabilità che aveva imparato, facendo la ragionevole ipotesi che i due alleli avessero la stessa probabilità $\frac{1}{2}$ e che il tutto avvenisse in condizioni di indipendenza.

Il discendente prende da ciascun genitore un allele, formando così una coppia che determinera il suo *genotipo* (cioè l'appartenenza a uno dei tre gruppi AA, Aa o aa) e il suo *fenotipo*, cioè la sua apparenza esteriore.

Lo schema che Mendel immaginò doveva essere di questo tipo:

$$\begin{array}{ccc} & A & a \\ A & AA & Aa \\ a & Aa & aa, \end{array}$$

con ciascuno dei quattro tipi di discendenti aventi la stessa probabilitị $\frac{1}{4}$. Naturalmente il genotipo Aa si presenta due volte.

Questo schema si accorda con i risultati sperimentali, perché la probabilità che sia presente un allele dominante (e quindi il suo fenotipo sia $\mathcal{A}$) è di $\frac{3}{4}$, mentre l'omozigote aa ha probabilità $\frac{1}{4}$. E quindi il rapporto tra i due fenotipi è di 3:1.

Gli alleli sono indistinguibili ed anche gli strumenti oggi disponibili non riescono a dire da quale genitore un individuo Aa(=aA) ha preso l'allele A (o a).

Mendel ebbe tantissime conferme sperimentali delle sue intuizioni. Per i suoi esperimenti usò circa 28.000 piante di piselli.

Allora non esisteva una strumentazione scientifica che potesse confermare le sue intuizioni: egli costruì però un modello, basato in parte sulla probabilità e in parte su congetture genetiche, che funzionava e che spiegava anche altri esperimenti che fece.

Ci dedicheremo ora ad un suo esperimento più complesso che doveva servire a confermare le sue intuizioni. Riguardava i semi, che potevano essere rotondi o grinzosi, gialli o verdi.

La prima delle due caratteristiche è dominante, la seconda recessiva.

Analizziamo intanto che cosa prevedeva il suo modello, che prevedeva anche che nella trasmissione delle due caratteristiche ci fosse indipendenza.

Indichiamo con A e a e B e b le due caratteristiche, dominante e recessiva, rispettivamente.

Lui ha incrociato due piante che erano bi-eterozigote, cioè del tipo AaBb.

Al discendente ogni genitore può "passare" quattro tipi di coppie di alleli: AB, Ab, aB e bb, ciascuna con probabilità $\frac{1}{4}$ (per l'indipendenza) e quindi, tenuto conto del contributo di entrambi i genitori, le possibilità sono descritte nello schema seguente:

	AB	Ab	aB	ab
AB	$AABB$	$AABb$	$AaBB$	$AaBb$
Ab	$AABb$	$AAbb$	$AaBb$	$Aabb$
aB	$AaBB$	$AaBb$	$aaBB$	$aaBb$
ab	$AaBb$	$Aabb$	$aaBb$	$aabb$

Ciascuna delle 16 combinazioni è altrettanto probabile. Mendel aveva la possibilità di scoprire il genotipo delle piante osservando la discendenza. Ma una prima indicazione si può ottenere già con l'osservazione del fenotipo, che è immediata.

La prima colonna e le quarta riga (quella più in alto) danno luogo allo stesso fenotipo (caselle $(1,1)$, $(1,2)$, $(1,3)$, $(1,4)$, $(2,4)$, $(3,4)$ e $(4,4)$). A queste sette tipologie vanno aggiunte ancora le due che si trovano nelle caselle $(2,2)$ e $(3,3)$: gli individui di queste nove classi posseggono tutti almeno un

allele A ed almeno un allele B. Dunque sono visibili le due caratteristiche nella versione dominante: semi rotondi e gialli.

Gli individui delle caselle $(2,1)$, $(2,3)$ e $(4,3)$ posseggono almeno un allele A, ma gli alleli dell'altra caratteristica sono in tutti e tre i casi del tipo bb. Semi rotondi e e verdi.

Una situazione simmetrica si presenta naturalmente con lo scambio delle due caratteristiche: gli individui delle caselle $(3,1)$, $(3,2)$ e $(4,2)$ posseggono almeno un allele B, ma rispetto all'altra caratteristica sono omozigoti con la coppia aa. Semi grinzosi e gialli.

Rimane fuori da questa classificazione solamente la casella $(4,1)$ che contiene individui che sono omozigoti rispetto ad entrambe le caratteristiche. Semi grinzosi e verdi.

Riassumendo: i quattro possibili fenotipi si presentano con probabilità $\frac{9}{16}$, $\frac{3}{16}$ $\frac{3}{16}$ e $\frac{1}{16}$ che stanno nel rapporto 9:3:3:1.

Nel 1865 Mendel fece due conferenze davanti ad un uditorio composto da chimici, biologi, botanici e medici, ma nessuno fu in grado, probabilmente, di capire l'importanza delle sue scoperte. L'anno dopo pubblicò su una rivista scientifica i suoi risultati e mandò una quarantina di copie del suo lavoro ai più famosi scienziati dell'epoca. La maggior parte non risposero, qualcuno addirittura lo scoraggiò di continuare le sue ricerche.

La pubblicazione di cui si parla ancora oggi ottenne solamente tre citazioni!

Nel 1900 tre scienziati, Hugo de Vries, Carl Correns e Erich von Tschermak pubblicarono indipendentemente e quasi contemporaneamente degli articoli che riprendevano le sue idee.

Rimasero molto delusi, immagino, quando scoprirono che erano stati preceduti da ricerche fatte trentacinque anni prima.

I risultati di quell'esperimento storico si sono conservati. Su un campione di 556 piante Mendel ottenne i seguenti risultati:

315 piante con i semi rotondi e gialli
101 piante con semi grinzosi e gialli
108 piante con semi rotondi e verdi
32 piante con semi grinzosi e verdi.

Ecco le frequenze relative ottenute nell'esperimento confrontate con le probabilità teoriche:

0.5665	0.5662
0.1816	0.1875
0.1942	0.1875
0.0575	0.0625

Molto convincenti, forse troppo.

Quando, agli inizi del '900, la statistica sviluppò strumenti di indagine più sofisticati, i suoi risultati furono sottoposti ad una critica "a posteriori".

Qualcuno (tra questi il grande statistico Fisher) avanzò l'ipotesi che Mendel avesse cercato di "abbellire" un po' i suoi risultati.

Per l'esempio di Mendel relativo ad un solo carattere, dove si trova un rapporto di 3:1 tra i due fenotipi, esisteva gà ai suoi tempi un metodo per scoprire se i dati osservati erano plausibili o no.

Non sappiamo però se lui conosceva tanta probabilità e statistica da saperli usare.

Una prima risposta la possiamo ricavare dall'approssimazione normale.

Nei suoi esperimenti Mendel usava un numero notevole di piante. Mai cifre tonde (e questo è credibile, lui usava quelle che c'erano), ma considerevoli. Per quello che si trova in letteratura, dai 500-600 a qualche migliaio.

Di conseguenza si può usare con stime abbastanza buone l'approssimazione normale.

L'evento che consideriamo è la comparsa dell'allele dominante. Potremmo ad esempio utilizzare l'esempio di prima, limitandoci a considerare il carattere che riguarda la forma del seme.

Su 556 piante, sono risultate $315 + 108 = 423$ piante del fenotipo rotondo e $101 + 32 = 132$ del fenotipo grinzoso. Risulta $\frac{423}{556} = 0.76$, molto vicino alle aspettative di Mendel. L'approssimazione normale ci consente di stimare la probabilità che questo avvenga e ci dice che questo non è impossibile, ma abbastanza improbabile.

Il ripetersi troppo frequente di queste vicinanze crea qualche dubbio sulla correttezza delle cifre fornite da Mendel.

Ma vedremo nel prossimo paragrafo un metodo più preciso per efettuare stime di questo genere.

Ma prima di affrontare questo argomento, vediamo ancora due esempi.

Il primo l'ho trovato in un libro di probabilità.

Su un campione di 200 famiglie con quattro figli era stata fatta un'indagine sul numero di maschi e femmine. Il risultato era il seguente:

0	maschi	26
1	maschio	49
2	maschi	54
3	maschi	45
4	maschi	26

Sappiamo che possiamo supporre che la nascita di un maschio o di una femmina ha circa la stessa probabilità e quindi è naturale pensare che il modello probabilistico appropriato sia una v.a. $B(4, \frac{1}{2})$. Ma questa v.a. darebbe, come frequenze attese, 12.5, 50, 75, 50, 12.5, numeri piuttosto lontani da quelli osservati.

È naturale porci la domanda se il modello probabilistico è adeguato oppure no.

Una spiegazione potrebbe essere un fenomeno del qual ho letto tempo fa: secondo quell'articolo nel genere umano il rapporto dei sessi dei nuovi nati sarebbe condizionato da un effetto di tipo sociale.

Per tutti i mammiferi esiste un meccanismo biologico che regola il rapporto tra i generi della discendenza in base ad uno stesso meccanismo.

Una delle teorie è che gli spermatozoi che sono portatori del cromosoma X e che fanno nascere discendenti maschi, siano più veloci, mentre gli altri siano più longevi.

Comunque sia, questo meccanismo può essere influenzato da fattori evolutivi e sociali.

Certi tipi di mammiferi hanno una struttura "sociale" che prevede una sovrabbondanza di femmine.

Per la specie umana è utile una leggera prevalenza di giovani maschi per i rischi legati alla caccia e alla difesa della tribù (più tardi le guerre e il lavoro.

È stato dimostrato ad esempio che come le due guerre mondiali vi sia stato un leggero, ma significativo, aumento di figli maschi.

Naturalmente non si può escludere che un meccanismo del genere funzioni non solo in rapporto alla società, ma anche all'interno di una famiglia. E non solo con un aumento dei figli maschi, se sono già nate delle femmine, ma anche con l'effetto simmetrico favorevole alla nascita di un maschio, se ci sono già in famiglia delle femmine.

Un altro esempio su cui si possono fare considerazioni utili è fornito dalla legge di Benford, che viene regolarmente utilizzata dall'ufficio delle tasse americano.

La legge di Benford

Incominciamo con un esperimento che ciascuno può ripetere per conto proprio.

Il lettore (che supponiamo si chiami Giovanni Rossi) prenda in mano l'elenco telefonico della propria città e lo apra alla pagina dove c'è il suo numero telefonico. Si annoti gli indirizzi (i numeri telefonici non ci interessano) che compaiono sulla stessa pagina.

Troverà, diciamo,

...
Rosselli Giuseppe, corso Cavour, 117
Rossi Francesco, via Crispi 23
Rossi Giovanni, via Dante 28
Rossi Roberto, via Petrarca 12
Rota Antonio, piazza della Repubblica, 7
...

e così via.

Osserviamo questi (immaginari) numeri civici e troviamo che essi iniziano con l'1 due volte (il 117 e il 12), con il 2 due volte (il 23 e il 28) e con il 7 una volta sola. Naturalmente il campione è striminzito e da esso non possiamo ricavare nessuna indicazione. Ma l'esperimento lo può fare ogni lettore su una scala un po' più significativa.

Su una normale pagina di un elenco telefonico priva di pubblicità ci sono circa 300 nomi ed indirizzi, quindi se uno vuole fare questo esperimento perderà un po' di tempo.

È abbastanza naturale attendersi di trovare i numeri civici che iniziano con l'1, il 2 e così via, fino al 9, equamente ripartiti, a parte le naturali fluttuazioni statistiche. Quindi in media ci aspettiamo di trovare ciascun numero al primo posto circa 30-35 volte. E invece chi proverà a fare l'esperimento (non importa se vive a Milano, Roma o Cosenza, se si chiama Rossi o Bianchi) troverà, molto probabilmente, che l'1 compare al primo posto circa il 30% delle volte, il 2 circa il 17% delle volte e così decrescendo, fino al 9 che sarà probabilmente il numero meno frequente, con un 5% di presenze al primo posto.

Alla fine dell'ottocento l'astronomo e matematico americano-canadese Simon Newcomb (1835–1909) osservo un fatto curioso, consultando un volume di tavole dei logaritmi.

Queste tavole erano usate nelle scuole fino all'avvento delle calcolatrici (ed anche un po' oltre, perché la scuola tarda ad adeguarsi). Esse consentivano di eseguire più rapidamente dei calcoli complicati, sia pure in maniera approssimata, perché con i logaritmi si possono trasformare prodotti in somme e le potenze in prodotti.

Newcomb, che nella sua carriera di astronomo aveva bisogno di far tanti conti, si accorse che le prime pagine delle tavole che usava erano più consumate delle ultime. In altre parole, capitava a lui e agli altri utilizzatori di quel volume, di dover calcolare più spesso il logaritmo di un numero che iniziava con l'uno o il due, piuttosto che di un numero che iniziava con l'otto o il nove.

Nel 1881 Newcomb scrisse un articolo al riguardo fornendo vari esempi, ma senza riuscire a dare una spiegazione del fenomeno osservato. Egli congetturò che la distribuzione di probabilità delle nove cifre fosse

$$p_k = \log_{10} \frac{k+1}{k}\,,$$

per $k = 1, 2, \ldots, 9$.

Questa è una densità di probabilità, perché

$$\sum_{k=1}^{9} \log_{10} \frac{k+1}{k} = \log_{10} \prod_{k=1}^{9} \frac{k+1}{k} = 1\,.$$

Siccome $\log_{10} 2 \simeq 0,30103$, da questa legge ricaviamo, ad esempio, che il numero 1 compare circa il 30% delle volte come prima cifra significativa.

Come qualche volta capita (ricordiamoci di Mendel), l'articolo non ebbe seguito e cadde nel dimenticatoio.

Passò più di mezzo secolo prima che il fisico ed ingegnere elettrotecnico americano Frank Benford (1883–1948) facesse la stessa osservazione. Benford nel 1938 pubblicò una lista di 20.229 numeri riproponendo la stessa distribuzione di probabilità. Benford fu più fortunato di Newcomb, l'articolo fu notato e la legge prese il suo nome.

Comunque anche l'articolo di Benford rimase senza una risposta da parte del mondo scientifico e solamente trent'anni dopo iniziarono ad apparire ricerche sulla "sua" legge. Ora esistono sull'argomento almeno tre libri e oltre cinquecento articoli.

La giustificazione della legge di Benford più convincente (anche se non spiega la sue presenza in tanti campi della vita quotidiana, ad esempio nell'elenco telefonico) è quella dell'*invaranza per scala*.

Consideriamo un grande insieme di dati statistici che derivano dalla misurazione di qualche grandezza: potremmo pensare ad esempio al consumo di elettricità o di gas degli utenti della provincia di Milano, oppure alla produzione di materie prime dei vari paesi del mondo.

Se esiste una distribuzione "naturale" per la prima cifra significativa di questi insiemi di dati, essa non dovrebbe dipendere dall'unità di misura che si usa (litri o galloni, tonnellate o barili, chilometri o miglia). Quindi, moltiplicando questo insieme di dati per una stessa costante, la distribuzione non dovrebbe cambiare.

Ralf Raimi, che scrisse uno dei primi articoli scientifici sulla legge di Benford, fu anche uno dei primi a utilizzare l'invarianza per scala per spiegare perché certi insiemi di dati verificano questa legge. Nell'introduzione del suo articolo scrisse:

"Se una tavola in cui sono segnate le superfici di un insieme di stati o di laghi è riscritta usando un'altra unità di misura, acri invece di ettari ad esempio, il risultato è un'altra tavola, nella quale tutti i numeri sono lo stesso multiplo del corrispondente numero della tavola originale. Se le prime cifre di tutte le tavole dell'universo hanno una fissata distribuzione di probabilità, questa distribuzione deve essere sicuramente indipendente dall'unità di misura scelta, poiché non risulta che Dio preferisca il sistema metrico a quello Inglese."

Vedremo ora, con delle considerazioni molto elementari, che se prendiamo per buono questo punto di vista, la distribuzione della prima cifra significativa non può essere uniforme.

Supponiamo di avere un elenco di distanze espresse in yarde e riscriviamo, in un altra tabella, le stesse distanze espresse in piedi. Denotiamo con p'_k e p''_k, per $1 \leq k \leq 9$, le frequenze delle prime cifre significative della prima e della seconda tabella, rispettivamente. E ricordiamo che tre piedi fanno una yarda.

Ogni numero della prima tabella che ha al primo posto l'1, è espresso nella seconda tabella con un numero che ha al primo posto 3, 4 o 5. La misura di 1,2 yarde equivale a 3,6 piedi, 1,4 yarde equivalgono a 4.2 piedi, 1,7 yarde sono 5,1 piedi e così via. Naturalmente è vero anche il viceversa: ogni misura in piedi che incominci per 3, 4 o 5, viene espressa nella prima tabella, in yarde, con un numero che inizia con l'1. Si ha dunque

$$p'_1 = p''_3 + p''_4 + p''_5 \,.$$

Considerazioni analoghe possono essere fatte con le misure che, in yarde, iniziano con un 2. Le corrispondenti misure espresse in piedi iniziano con 6, 7 o 8. Si ottiene così una seconda equazione

$$p'_2 = p''_6 + p''_7 + p''_8 \,.$$

Ogni regolarità statistica presente nella prima tabella deve essere presente anche nella seconda e in ogni altra tabella ottenuta dalla prima usando altre unità del fantasioso sistema "imperiale britannico", come line, inch (pollice), fathom, rod, chain, furlong o miglia. Il sistema metrico naturalmente non offre spunti altrettanto interessanti.

Se assumiamo che le distribuzioni p_k, $1 \leq k \leq 9$, nelle due tabelle hanno la stessa distribuzione, dobbiamo concludere che

$$p_1 = p_3 + p_4 + p_5$$

e

$$p_2 = p_6 + p_7 + p_8 \,.$$

Altre equazioni si possono ottenere confrontando tabelle relative ad altre unità di misura. Potremmo ad esempio prendere le yarde e i fathom (1 fathom = 2 yarde).

Procediamo come abbiamo fatto prima e vediamo che cosa si trova.

Se un tavolo è lungo tra un fathom e $1,99$ fathom, la corrispondente misura in yarde sarà compresa tra 2 e $3,98$. Ne deduciamo quindi che

$$p_1 = p_2 + p_3 \,,$$

e procedendo analogamente con le misure, in fathom, comprese tra 2 e 3, tra 3 e 4 e tra 4 e 5, otteniamo complessivamente altre tre equazioni

$$p_2 = p_4 + p_5 \,,$$

$$p_3 = p_6 + p_7 \,,$$

e

$$p_4 = p_8 + p_9 \,.$$

Visto che i numeri p_i, $1 \leq i \leq 9$ sono una distribuzione di probabilità, c'è un altra ovvia equazione che lega i nove numeri e cioè

$$\sum_{k=1}^{9} p_k = 1 \,.$$

Abbiamo trovato così sette equazioni lineari che sono linearmente indipendenti tra loro. Si potrebbe pensare di procedere in questa maniera con altre coppie di unità di misura il cui rapporto è un intero, ma sembra che questo procedimento non porti oltre.

Oltre ai rapporti 1:2 e 1:3 che abbiamo appena utilizzato, resterebbero solo altre due possibilità: il rapporto 1:4 e 1:5.

La prima delle due possibilità ci porta all'altra coppia di misure britanniche, chiamate chain e rod (1 chain = 4 rod). L'argomentazione analoga a quella vista prima ci darebbe l'equazione

$$p_1 = p_4 + p_5 + p_6 + p_7 \,.$$

Questa equazione è diversa dalle altre, ma è deducibile dalle precedenti, poiché la prima equazione ci dà $p_1 = p_3 + p_4 + p_5$, mentre la quinta ci dice che $p_3 = p_6 + p_7$. Sostituendo la seconda nella prima, troviamo la "nuova" equazione.

Anche l'ultima delle possibilità non ci porta da nessuna parte. Inventiamoci noi due misure che siamo in rapporto 1:5, chiamiamole braccio e pertica (ci sono due antiche misure italiane che sono circa in questo rapporto) e ripetiamo il ragionamento visto prima.

L'equazione che se ne ricava è

$$p_1 = p_5 + p_6 + p_7 + p_8 + p_9 \,.$$

Ma anche questa, come si vede facilmente, si ricava dalle altre. Infatti

$$p_1 = p_3 + p_4 + p_5 = p_6 + p_7 + p_4 + p_5 = p_6 + p_7 + p_8 + p_9 + p_5 \,.$$

Notiamo che le probabilità π_k della legge di Benford, cioè

$$\pi_k = \log_{10} \frac{k+1}{k} \,,$$

con $1 \leq k \leq 9$, sono una soluzione particolare del sistema delle sette equazioni. Il sistema però ammette infinite soluzioni (∞^2, scrive qualcuno).

D'altra parte, mentre è possibile che ci sia un'ottava equazione lineare a coefficienti interi, indipendente dalle precedenti (che chi scrive non è riuscito a scoprire), sicuramente non ne esiste una nona (sempre con coefficienti interi), perché (come è stato dimostrato negli articoli citati nella bibliografia) solamente la distribuzione di Benford è compatibile con l'invarianza per scala, e i π_k sono tutti irrazionali, mentre la regola di Cramer (o il metodo dell'eliminazione successiva) di un sistema ottenuto come descritto sopra darebbe una soluzione razionale.

Dimostriamo ora che la distribuzione uniforme $q_i = \frac{1}{9}$ non è una soluzione del sistema. Questo seguirà dalla seguente proposizione che è un po' più generale.

Proposizione
Se r_i, $1 \leq i \leq 9$ è una soluzione del sistema lineare delle sette equazioni trovate, e se $r_i > 0$ per ogni i, allora $r_1 > r_j$, per ogni $j \geq 2$.
Dimostrazione. Siccome $r_1 = r_2 + r_3$, si ha che $r_1 > r_2$ e $r_1 > r_3$. Da $r_1 = r_3 + r_4 + r_5$, segue anche che $r_1 > r_4$ e $r_1 > r_5$. Risulta d'altra parte $r_2 = r_6 + r_7 + r_8$, da cui si deduce che $r_2 > r_6$, $r_2 > r_7$ ed $r_2 > r_8$ e poiché $r_1 > r_2$, si ha anche che $p_1 > p_6, p_7$ e p_8. Ma da $r_4 = r_8 + r_9$ segue che $(r_1 >) r_4 > r_9$ e quindi la tesi.

La legge di Benford può essere presentata a degli studenti interessati che frequentano un liceo e si può tentare anche di dimostrare la proposizione precedente.

Ma noi andremo oltre.

Ogni numero reale positivo x può essere scritto in quella che alle volte si chiama notazione scientifica, come

$$x = a.b_1 b_2 \ldots \ 10^c,$$

dove a è un intero $1 \leq a \leq 9$, i b_i sono degli interi tali che $0 \leq b_i \leq 9$ per ogni i, mentre $c \in \mathbb{Z}$. Per evitare ambiguità, escludiamo le rappresentazioni nelle quali $b_i = 9$ per tutti gli i sufficientemente grandi.

Definiamo, per $x \in \mathbb{R}_+$

$$M(x) = a.b_1 b_2 \ldots$$

e

$$S(x) = a \ .$$

L'applicazione M manda $\mathbb{R}_+$ su $[1, 10)$ mentre S applica $\mathbb{R}_+$ su $\{1, 2, \ldots, 9\}$. Il numero $S(x)$ è detto *la prima cifra significativa* di x, mentre M potremmo chiamarla *mantissa*.

Si osservi che $M(x) = M(y)$ se e solo se $x = y \ 10^c$ per qualche intero c. Inoltre possiamo definire sui numeri reali positivi la seguente relazione di equivalenza:

xMy se e solo se $M(x) = M(y)$. L'insieme $[1, 10)$ rappresenta tutte le classi di equivalenza.

L'insieme $[1, 10)$ è quindi lo spazio campione che contiene la prima cifra significativa di tutti i numeri reali positivi. Su questo insieme consideriamo la σ-algebra di Borel.

Per ogni boreliano $B \subset [1, 10)$ e per ogni $c > 0$, definiamo il prodotto cB nel modo seguente:

$$cB = \{y : y = cx \bmod [1, 10), x \in B\}.$$

Definiamo ora l'invarianza per scala di una v.a X che assume i suoi valori in $[1, 10)$.

Diciamo che una v.a X che prende i valori in $[1, 10)$ è invariante per scala, se per tutti gli $x, y \in [1, 10)$ si ha

$$P(1 \le X < x) = P(y \le X < xy),$$

se $xy < 10$, mentre se $xy \ge 10$,

$$P(1 \le X < x) = P(y \le X < 10) + P(1 \le X < xy10^{-1}).$$

(In questo caso l'intervallo $[y, xy[$ resta diviso in due. $[y, 10[$ è ovviamente contenuto in $[1, 10)$, mentre $[10, xy)$ è contenuto in $[10, 100)$. I numeri in esso contenuti hanno la mantissa che appartiene all'intervallo $[1, xy10^{-1}[$).

Teorema

Una v.a. con valori in $[1, 10)$ è invariante per scala se e solo se la sua funzione di ripartizione F (nulla per $x < 0$ e identicamente uguale a 1 per $x \ge 10$) è definita per $x \in [1, 10]$,

$$F(x) = \log_{10} x.$$

Dimostrazione. Supponiamo che la funzione di ripartizione sia quella appena descritta. Ovviamente la X prende valori in $[1, 10)$. Verifichiamo che è invariante per scala.

Sia $x \in [1, 10)$ e supponiamo che $y \in [1, 10)$ sia tale che $xy < 10$. Valutiamo

$$P(y \le X < xy) = F(xy) - F(y) = \log_{10} xy - \log_{10} y = \log_{10} x = P(1 \le X < x).$$

Se invece $xy \ge 10$, dobbiamo valutare

$$P(y \le X < 10) + P(1 \le X < xy\,10^{-1}) = F(10) - F(y) + F(xy\,10^{-1}) - F(1) =$$

$$\log_{10} 10 - \log_{10} y + \log_{10}(xy10^{-1}) = \log_{10} x = P(1 \le X < x).$$

Dimostriamo ora il viceversa.

Supponiamo che X assuma i suoi valori in $[1, 10)$ e sia invariante per scala. Dobbiamo dimostrare che la sua funzione di ripartizione è quella descritta nell'enunciato del teorema. Se $xy < 10$, l'invarianza per scala può essere riscritta in termini della sua funzione di ripartizione F come

$$F(x) + F(y) = F(xy)\,.$$

Ponendo $x = 10^u$ e $y = 10^v$, si ha che $u, v \in [0, 1)$, $uv < 1$ e se noi definiamo $g(u) = F(10^u)$, otteniamo per g l'equazione di Cauchy

$$g(u) + g(v) = g(u + v)\,.$$

Si osservi che g e la composta di due funzioni non decrescenti, quindi la g stessa è non decrescente. Inoltre $g(0) = 0$ e $g(1) = 1$, poiché $F(1) = 0$ e $F(10) = 1$.

È ben noto che le soluzioni monotone dell'equazione funzionale di Cauchy sono lineari. Ma in questo caso dobbiamo tener conto delle restrizioni su u e v. Verificheremo che pur con queste restrizioni la conclusione rimane la stessa.

Dall'equazione trovata per la g vediamo che $g(2u) = g(u + u) = g(u) + g(u) = 2g(u)$ per tutti gli $u \in [0, \frac{1}{2}]$. Per induzione vediamo che per ogni intero positivo n

$$g(nu) = ng(u)$$

per ogni $u \in [0, \frac{1}{n}]$. Per ogni $v \in [0, 1]$, applichiamo quest'ultima equazione a $u = \frac{v}{n}$ ottenendo

$$g\left(\frac{v}{n}\right) = \frac{1}{n}g(v)$$

per ogni $v \in [0, 1]$. Siano ora m e n interi positivi tali che $m \leq n$ e sia $u \in [0, 1]$. Applicando le equazioni appena ottenute si ha che

$$g\left(\frac{m}{n}u\right) = mg\left(\frac{u}{n}\right) = \frac{m}{n}g(u)\,.$$

Da $g(1) = 1$ segue immediatamente che $g(q) = q$ per tutti i numeri razionali $u \in [0, 1]$. Dalla monotonia di g segue che $g(u) = u$ per ogni $u \in [0, 1]$. Sostituendo ora $u = \log_{10} x$ in $g(u) = F(10^u)$, vediamo che

$$F(x) = \log_{10} x$$

per tutti gli $x \in [1, 10]$.

Dal teorema precedente possiamo immediatamente dedurre la seguente conseguenza.

Corollario

Una v.a X a valori in $[1, 10)$ è invariante per scala se e solo se X è assolutamente continua e la sua densità è

$$f(x) = \frac{\log_{10} e}{x}$$

per $x \in [1, 10]$ e 0 altrove.

Vediamo ora che la v.a appena descritta verifica la legge di Benford.
Dal teorema precedente si ha che

$$p_k = P(S(x) = k) = P(k \leq X < k + 1) = F(k + 1) - F(k) = \log_{10} \frac{k + 1}{k} .$$

Legge di Benford ed evasione fiscale

Dagli inizi degli anni settanta l'ufficio delle tasse americano ha iniziato ad utilizzare la legge di Benford per scoprire possibili evasori fiscali.

Presumendo che chi si inventa dei numeri (e presenta false fatture, spese mediche fasulle, assicurazioni ed altri oneri deducibili inventati) tenda a equidistribure le cifre, violando la legge di Benford.

Degli studiosi americani analizzarono delle statistiche macroeconomiche come ad esempio i bilanci degli stati e trovarono che la maggioranza di questi obbedivano alla legge di Benford, mentre alcuni stati con un'economia meno trasparente, la violavano.

Gli esperimenti di Mendel, l'esempio del numero dei figli maschi e la legge di Benford presentano una caratteristica comune e nuova.

Non si tratta più di stimare speranza matematica e (eventualmente) varianza da cui dipende la densità della v.a. su cui stiamo indagando, ma di valutare se sia accettabile un insieme di frequenze associate ad una v.a. incognita che assume un numero finito di m valori che possiamo identificare con gli interi $\{1, 2, \ldots, m\}$.

Vediamo alcuni esempi (e vediamo anche come si inseriscano in questo quadro gli esempi che abbiamo appena descritto).

Un primo esempio potrebbe essere un dado che non sappiamo se sia regolare o meno. Potremmo effettuare ad esempio 600 lanci e ottenere per i punteggi da 1 a 6 le seguenti frequenze: 102, 86, 97, 104, 121, 90. Ci chiediamo se il dado può considerarsi regolare ad un certo livello di confidenza.

Già a questo semplice esempio non possiamo applicare nessuna delle tecniche utilizzate prima.

Riguardo agli esperimenti di Mendel, gli unici dati trattabili con i metodi studiati finora è sono quelli relativi alla presenza del fenotipo dominante di una sola caratteristica. Ma l'esperimento che dǫme rapporto tra i fenotipi nel rapporto 9:3:3:1 sfugge a questa possibilità.

Così è anche per l'esempio delle 200 famiglie con quattro figli, classificate per numero di figli maschi.

L'ultimo esempio trattato è quello della legge di Benford, dove i parametri osservati sono nove e si tratta di vedere se le osservazioni sono compatibili, ad un certo livello di confidenza, con tale legge.

Test del χ^2

Supponiamo che la v.a. sotto osservazione X sia discreta e che assuma m possibili valori (spesso sono proprio i numeri $\{1, 2, \ldots, m\}$).

Il parametro ϑ da stimare è un'm-upla $(\vartheta_1, \vartheta_2, \ldots, \vartheta_m)$ tale che $\sum_{i=1}^m \vartheta_i = 1$ e $\vartheta_i \geq 0$.

In questo paragrafo vogliamo analizzare la seguente situazione: sul fenomeno aleatorio che stiamo studiando facciamo una certa ipotesi $H_0 = (p_1, p_2, \ldots, p_m)$ e vogliamo vedere se i dati raccolti smentiscono l'ipotesi fatta.

Un tipico caso potrebbe essere il seguente: vogliamo verificare se un dado è regolare e sottoponiamo ad un'indagine l'ipotesi $H_0 = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$.

Facciamo n osservazioni ed indichiamo con

$$N_k^n \text{ e } \bar{p}_k^n = \frac{1}{n} N_k^n$$

le frequenze e le frequenze relative degli m valori, $1 \leq k \leq m$.

La legge dei grandi numeri ci dice che, al tendere di n all'infinito, $\bar{p}_k^n \to p_k$ (in probabilità o, meglio, quasi certamente), dove p_k sono i valori effettivi dei parametri della v.a. X.

Pearson (1857-1936) ha proposto la seguente statistica:

$$T_n = \sum_{k=1}^m \frac{1}{np_k} (N_k - np_k)^2 = n \sum_{k=1}^m \frac{(\bar{p}_k - p_k)^2}{p_k} \,.$$

Nella formula abbiamo omesso di scrivere l'apice n, intendendo che $N_k^n = N_k$ e $\bar{p}_k^n = \bar{p}_k$.

Teorema

Se le $X_1, X_2, \ldots, X_n$ sono copie indipendenti della stessa v.a. X la cui legge dipende dai parametri $\vartheta = (p_1, p_2, \ldots, p_m)$, allora la successione $\{T_n\}$ converge in legge alla v.a. $\chi^2(m-1)$.

Esempio 1

Supponiamo di lanciare un dado 2000 volte ottenendo i seguenti risultati

punteggio	1	2	3	4	5	6
n^0 successi	382	322	314	316	344	316

Le frequenze relative sono: $\bar{p}_1 = 0.194$, $\bar{p}_2 = 0.161$, $\bar{p}_3 = 0.157$, $\bar{p}_4 = 0.158$, $\bar{p}_5 = 0.172$ e $\bar{p}_6 = 0.158$. L'ipotesi di regolarità corrisponde a $p_k = 0.167$.

Una regola pratica perché si possa ritenere che la legge di T_n sia sufficientemente vicina alla $\chi^2(m-1)$ è che $np_k > 5$ per $1 \leq k \leq m$.

Nell'esempio questa regola è soddisfatta.

Se facciamo i calcoli, troviamo che

$$T_n = 2000 \sum_{k=1}^{6} 6(\bar{p}_k - 0.167)^2 = 12.6\,.$$

Al livello di confidenza del 95% ci sentiamo di poter respingere la regolarità del dado, poichè il valore che si osserva nella tabella del $\chi^2(5)$ è $11,07$. Al livello di confidenza del 99% corrisponde il valore tabellare 15.09 e quindi a quel livello di confidenza la regolarità del dado non può essere respinta.

Osservazione

Dalla seconda forma della T_n si vede facilmente che si moltiplica n per un coefficiente γ tale che $\gamma\, n \in \mathbb{N}$, allora

$$T_{\gamma n} = \gamma\, T_n\,.$$

Da questo segue che se nell'esempio precedente fossero state fatti solamente 1000 lanci e se le frequenze relative osservate fossero le stesse, allora il valore di T_n sarebbe risultato dimezzato, rendendo non significativi i dati raccolti.

Gli scostamenti dalle probabilità teoriche sarebbero gli stessi, ma il campione sarebbe troppo modesto per poter fare delle conclusioni sul dado osservato.

Esempio 2

Riconsideriamo l'esempio delle 200 famiglie con (almeno) quattro figli classificate a seconda del numero dei figli maschi.

L'ipotesi da sottoporre a verifica è se il genere dei figli che nascono ad una fissata coppia nel primo, secondo, terzo e quarto parto si possono considerare indipendenti oppure no. Dobbiamo quindi verificare se i dati raccolti sono compatibili con quelli che risulterebbero da una binomiale $B(4, \frac{1}{2})$. Questo se accettiamo, con una certa approssimazione, che la nascita di un maschio abbia probabilità $\frac{1}{2}$

(sappiamo che non è così e che la probabilità che nasca un maschio è pari circa a $0,52$).

I dati riportati nell'esempio erano i seguenti:

numero maschi	0	1	2	3	4
numero famiglie	26	49	54	45	26

Le frequenze empiriche sono dunque $\bar{p}_0 = 0,13$, $\bar{p}_1 = 0,245$, $\bar{p}_2 = 0,27$, $\bar{p}_3 = 0,225$ e $\bar{p}_4 = 0,13$. Quelle teoriche sono invece $p_k = \binom{4}{k}2^{-4}$. Anche in questo caso l'approssimazione della T_n con la $\chi^2(4)$ è sufficientemente buona, perché $200p_0 = 12,5 > 5$.

Si ha che $T_{200} = 35,56$, largamente superiore ai valori tabellari della $\chi^2(4)$.

Notiamo anzi che le famiglie censite hanno messo al mondo più femmine che maschi: 404 contro 396, quindi il confronto dei dati sperimentali con la $B(4,0,52)$ darebbe un valore della T_{200} ancora più grande.

Esempio 3

In base ai dati dell'esempio di Mendel si trova che T=.456, valore che si trova circa a metà tra i valori della $\chi^2(3)$ corrispondenti alle percentuali del 5% e del 10%.

Di per sé niente di drammatico. Un caso fortunato, ma che rientra tutto sommato nella normalità. La critica di Fisher si appuntava sul fatto che Mendel, nei tanti esperimenti che aveva fatto, era stato quasi sistematicamente fortunato.

Esempio 4

Supponiamo di voler verificare se dei dati raccolti sono compatibili con l'ipotesi che la v.a. X studiata sia di Poisson di parametro $\lambda = \frac{1}{2}$.

Il test di Pearson non sembra applicabile, perché la X assume infiniti valori, mentre noi abbiamo fatto l'ipotesi che i valori assunti sono in numero finito. Solo con questa ipotesi possiamo utilizzare la $\chi^2(m-1)$. Sulla tabella cercheremmo inutilmente la $\chi^2(\infty)$!

Osserviamo che, per il teorema centrale del limite, se $X_m \sim \chi^2(m)$, allora $\frac{X_m - m}{\sqrt{2m}}$ tende alla $N(0,1)$.

Ma c'è un altro modo per utilizzare il test del χ^2 nel caso in cui la v.a. sia discreta ed assuma infiniti valori, ed anche nel caso che la v.a. sia continua.

Si possono infatti suddividere i valori assunti dalla v.a. in m classi scelte con un criterio che di volta in volta ci può sembrare adeguato.

Nel nostro caso potremmo ad esempio suddividere i valori nelle classi $\{k = 0\}$, $\{k = 1\}$, $\{k = 2\}$, $\{k = 3\}$ e $\{k > 3\}$.

Le probabilità teoriche da sottoporre a verifica sono $p_k = e^{-\frac{1}{2}} \frac{1}{2^k k!}$ per $k = 0, 1, 2, 3$ e $p_4 = 1 - \sum_{k=0}^{3} p_k$. Usando la solita formula otteniamo

$$p_0 = 0,6065\,,$$

$$p_1 = 0,3032\,,$$

$$p_2 = 0,0758\,,$$

$$p_3 = 0,0125\,,$$

$$p_4 = 0,0219\,.$$

Supponiamo di aver fatto 100 esperimenti e di aver ottenuto 55 volte lo 0, 27 volte l'1, 5 volte il 2, 2 volte il 3 e 11 volte un numero superiore a 3. Già ad un'osservazione superficiale appare che c'è una presenza eccessiva di valori superiori a 3. Vediamo che cosa dice il test.

A conti fatti troviamo $T = 38,34$, superiore ad ogni valore tabellare della $\chi^2(4)$. In effetti già il solo ultimo addendo, quello relativo alla p_4, dà un contributo di più di 35.

Esempio 5

Supponiamo ora che la v.a. X da testare sia un'esponenziale e facciamo l'ipotesi che il parametro sia uguale a 1.

Supponiamo di aver osservato su 100 esperimenti 62 del tipo $\{X \in [0, 1[\}$, 23 del tipo $\{X \in [1, 2[\}$, 9 del tipo $\{X \in [2, 3[\}$, 4 del tipo $\{X \in [4, 5[\}$ e 2 del tipo $\{X \geq 5\}$.

Si noti che in questo caso il parametro è noto, quello che si vuole capire è se è ragionevole il modello, cioè la scelta della v.a. esponenziale.

Sappiamo che $P(X \in [k, k+1[) = e^{-k} - e^{-k-1} = e^{-k}(1 - e^{-1})$ per $k \geq 0$.

Le probabilità teoriche delle cinque classi sono $p_1 = 0,6321$, $p_2 = 0,2326$, $p_2 = 0,0856$, $p_3 = 0,0315$, $p_4 = 0,0116$ e $p_5 = 0,0076$.

Il calcolo ci dà

$$T = 20,29\,,$$

numero che supera tutti quelli presenti nella tabella per la $\chi^2(4)$, quindi l'ipotesi va rifiutata: con probabilità superiore al 99% i dati raccolti non corrispondono ad un'esponenziale di parametro 1.

Esempio 6

Vediamo un esempio (costruito a tavolino) di applicazione della legge di Benford.

Siccome le classi sono nove e sono anche molo diverse per dimensione, conviene accorpare alcune di esse.

Le probabilità di avere la prima cifra significativa uguale a k è $\log_{10} \frac{k+1}{k}$, ovvero

1	2	3	4	5	6	7	8	9
0.301	0.176	0.125	0.097	0.079	0.067	0.58	0.051	0.048

Si può, ad esempio, considerare le seguenti classi: $\{1\}$, $\{2\}$, $\{3\}$, $\{4,5\}$ e $\{6,7,8,9\}$ che hanno rispetto alla legge di Benford, probabilità $p_1 = 0.301$, $p_2 = 0.176$, $p_3 = 0.125$, $p_4 = 0.175$, $p_5 = 0.223$.

Supponiamo che un contribuente presenti ad accompagnamento della dichiarazione dei redditi 250 documenti. Le cifre riportare in essi iniziano con l'1 in 40 casi, con il 2 in 40 casi, con il 3 in 30 casi, con il 4 o il 5 in 50 casi e con il 6, il 7, l'8 o il 9 in 90 casi.

Le frequenze relative sono $f_1 = 0.16$, $f_2 = 0.16$, $f_3 = 0.12$, $f_4 = 0.2$, $f_5 = 0.36$.

Bisogna ora seguire tutta la trafila: calcolare i cinque numeri $(p_i - f_i)$, elevarli al quadrato, dividere ciascuno di questi cinque quadrati per i p_i, sommare i cinque numeri così ottenuti e infine moltiplicare questa somma per 250, il numero dei documenti.

Il risultato è 75, ben lontano da ogni valore tabulare della $\chi^2(4)$.

L'ufficio delle tasse americano metterebbe questo contribuente sotto la lente di ingrandimento come un sospetto evasore.

Certo, se poi viene fuori che la persona sospetta è un commerciante il cui business principale è la vendita di scope elettriche a 99.99 dollari l'una, quel contribuente non verrà mai a sapere di essere stato "indagato".

Esempio 7

Vediamo un altro esempio con la v.a. di Poisson, ma questa volta supponiamo di ignorare il suo parametro.

I dati sperimentali siano questi: una fabbrica ha prodotto 500 macchinari. Di questi, 225 sono privi di difetti, 183 hanno un difetto, 64 ne hanno 2, 23 ne hanno 3 e 5 ne hanno 4 o più.

Vorremmo sapere se questa distribuzione di difetti può seguire una legge di Poisson.

La media empirica dà una stima del parametro λ che, appunto, non conosciamo. Un facile calcolo ci dà $\bar{X} = 0.8$.

Assunto allora $\lambda = 0.8$, le probabilità con cui confrontare i dati sperimentali sono:

$$p_0 = 0.4493\,,$$

$$p_1 = 0.3594\,,$$
$$p_2 = 0.1438\,,$$
$$p_3 = 0.0383\,,$$
$$p_4 = 0.0082\,.$$

Invece le frequenze relative osservate sono

$$f_0 = 0.4493\,,$$
$$f_1 = 0.3594\,,$$
$$f_2 = 0.1438\,,$$
$$f_3 = 0.0383\,,$$
$$f_4 = 0.0082\,.$$

A questo punto riparte la macchinetta: Si calcolano le differenze $p_i - f_i$, se ne prendono i quadrati. si dividono ciascuno per il suo p_i, so sommano e la somma viene moltiplicata per 500. Si ottiene alla fine che $T = 10$.

Poiché $\chi^{0.95}(4) = 9.49$ e $\chi^{0.99}(4) = 13.28$, ci ritroviamo nuovamente in una di quelle situazioni in cui accettiamo il modello a un livello di confidenza del 95%, ma non lo accettiamo ad un livello di confidenza superiore del 99%.

Passiamo ora ad una variante del test di Pearson che si utilizza quando si vuole avere indicazioni sul fatto che due v.a. sono indipendenti oppure no.

Il test dell'indipendenza

La densità χ^2 è utile anche per fare delle verifiche statistiche sull'indipendenza di due o più caratteristiche. Supponiamo di avere una "popolazione" (intesa in senso statistico) e di rilevare in un campione due caratteristiche degli "individui" esaminati. Potremmo ad esempio, tra i laureati dell'università della Calabria, esaminare un campione casuale e rappresentativo e registrare la tipologia del titolo conseguito (ingegneria, economia, informatica, chimica, ecc) [o invece la durata degli studi, o il voto finale] e per le stesse persone lo status lavorativo a distanza di tre anni dalla laurea.

Oppure potremmo registrare contemporaneamente estrazione sociale [o genere, o area geografica] e preferenze politiche, oppure, per un gruppo di pazienti, le diverse cure seguite e loro guarigione, concentrazioni di due sostanze in vari campioni di acqua e così via.

Sappiamo che tra la densità congiunta e le densità marginali nel caso di variabili aleatorie discrete indipendenti c'è la relazione

$$p(x_i, y_j) = p_1(x_i)p_2(y_j)\,.$$

Ne segue che se due caratteristiche rilevate su un campione sono indipendenti tra loro, le frequenze relative dovrebbero, approssimativamente, verificare la relazione

$$f(x_i, y_j) = f_1(x_i)f_2(y_j)\,.$$

Per vedere se c'è indipendenza oppure no, l'idea è quella di confrontare i termini a sinistra e destra di questa equazione.

Prima di introdurre i simboli necessari per enunciare il teorema successivo, vediamo un esempio.

Nella tabella successiva sono riportati i risultati (immaginari) di un inchiesta fatta tra 1150 elettori americani. Essi sono suddivisi per preferenze politiche (D=democratici, R= repubblicani, I=indipendenti) e per genere (M=maschio, F=femmina).

	D	R	I	
M	300	200	100	600
F	350	150	50	550
	650	350	150	1150

Il significato della tabella è abbastanza chiaro: nella casella della colonna R e riga F è indicato il numero (150) delle donne che hanno votato per i repubblicani.

Ci possiamo porre il problema se i due caratteri rilevati sono indipendenti. Da una lettura superficiale sembrerebbe che le donne abbiano espresso, in percentuale, più preferenze ai democratici degli uomini. La differenza è statisticamente significativa?

In generale, supponiamo di avere un campione di rango N e che le due caratteristiche assumano r ed s valori, rispettivamente. Useremo gli indici i, per $1 \leq i \leq r$ e, rispettivamente, j, per $1 \leq j \leq s$ per individuare le modalità con cui si verificano le caratteristiche su cui raccogliamo i dati.

Indichiamo con $N_{i,j}$ il numero degli elementi del campione che occupano la casella (i, j). Indicheremo inoltre con N_i il numero degli elementi dell'i-esima riga e con N^j gli elementi della j-esima colonna.

Illustriamo queste notazioni con riferimento all'esempio precedente:

	X_1	X_2	X_3	
Y_1	N_{11}	N_{12}	N_{13}	N_1
Y_2	N_{21}	N_{22}	N_{23}	N_2
	N^1	N^2	N^2	N

L'idea principale che si utilizza per verificare l'indipendenza delle due v.a. X e Y che possono, nel nostro esempio, assumere tre e due valori, rispettivamente, consiste nel confrontare, per ogni coppia di indici (i, j) la frequenza relativa osservata

$$\frac{N_{ij}}{N}$$

degli individui del campione che hanno entrambe le caratteristiche i e j e il prodotto delle frequenze marginali osservate e cioè le frequenze degli individui che hanno le caratteristiche i e j solamente, in rapporto a tutto il campione, ovvero

$$\frac{N_i}{N}\frac{N^j}{N}\,.$$

Anche per questo test si usano i quantili della v.a. χ^2.

Conviene introdurre le seguenti notazioni: $\bar{\pi}_{ij} = \frac{N_{ij}}{N}$, $\bar{p}_i = \frac{N_i}{N}$ e $\bar{q}_j = \frac{N^j}{N}$.

Vale allora il seguente teorema.

Teorema

La variabile aleatoria

$$T = n \sum_{i=1}^{m} \sum_{j=1}^{r} \frac{(\bar{p}_i \bar{q}_j - \bar{\pi}_{ij})^2}{\bar{\pi}_{ij}}$$

converge in legge ad una v.a. di legge $\chi^2((m-1)(r-1))$.

Questo risultato dà un criterio per accettare o respingere l'ipotesi che due v.a. siano indipendenti.

Esempio

Per controllare l'efficacia del vaccino contro la poliomielite, creato dal dottor Jonas Salk (e successivamente sostituito dal più efficace vaccino dovuto ad Albert Sabin) è stato fatto un test.

Vennero scelti a caso due gruppi omogenei di bambini, uno dei quali è stato sottoposto alla vaccinazione, mentre il secondo gruppo doveva servire da *gruppo di controllo*. I due gruppi erano molto numerosi, ciascuno composto da oltre 200.000 bambini.

A distanza un tempo prefissato i due gruppi vennero controllati e suddivisi in tre categorie: del primo gruppo facevano parte coloro che non si erano ammalati, del secondo, coloro che si erano ammalati ma che non avevano subito conseguenze (alcuni casi erano rimasti asintomatici), nel terzo coloro che hanno subito una paralisi (abitualmente degli arti inferiori).

Ecco i dati

	primo gruppo	secondo gruppo	terzo gruppo	
vaccinati	200.688	24	33	200.745
non vaccinati	201.087	27	115	201.229
	400.775	51	148	401.974

Per verificare se i dati supportano, o meno, l'efficacia del vaccino, occorre fare diversi calcoli.

La tabella successiva riportata i valori $\bar{\pi}_{ij}$

	primo gruppo	secondo gruppo	terzo gruppo
vaccinati	0.49926	$5.97\,10^{-5}$	$8.20\ 10^{-5}$
non vaccinati	0.50025	$6.62\,10^{-5}$	$28.60\,10^{-5}$

Il passo successivo consiste nel calcolo dei prodotti $\bar{p}_i\bar{q}_j$.

	primo gruppo	secondo gruppo	terzo gruppo
vaccinati	0.49915	$6.33\,10^{-5}$	$18.38\ 10^{-5}$
non vaccinati	0.50025	$6.35\,10^{-5}$	$18.43\,10^{-5}$

Il valore dello stimatore calcolato con la formula che abbiamo visto dà 45.42 che è molto più grande de valore tabulare $\chi^2_{0.95}(2) = 5.99$, ma anche del valore $\chi^2_{0.99}(2) = 9.21$ e quindi l'indipendenza è respinta con largo margine.

I dati riportati sono piuttosto datati e sicuramente esistono ricerche più recenti sui nuovi vaccini che sono stati sviluppati nel corso degli anni.

I primi vaccini sono stati testati nel lontano 1950.

Quello di Salk è stato messo in vendita nel 1955, quello di Sabin nel 1961.

Ma successivamente ci sono stati altri progressi e i vaccini attuali sono molto efficaci, tanto da estirpare la malattia da tutto il mondo sviluppato.

I due vaccini hanno praticamente eliminato la poliomielite dal mondo riducendo il numero dei casi registrati ogni anno da 350.000 nel 1988 a 33 nel 2018.

Negli Stati Uniti prima e in Europa dal 2000, la poliomielite è stata completamente eradicata.

Permane solamente in regioni nelle quali le condizioni sanitarie sono precarie e altre regioni nelle quali ci sono pregiudizi religiosi contro le vaccinazioni in genere.

Il vaccino orale causa circa tre casi di poliomielite paralitica su un milione di vaccinazioni. Questo numero va comparato con i 1000 casi su un milione di persone che restano paralizzate non essendo coperte dal vaccino.

Coloro che sono contrari alla vaccinazione (contro la poliomielite e contro altre malattie invettive) leggono il primo numero, ma non il secondo e non si rendono conto che il vaiolo è stato completamente estirpato dal mondo (inclusi i paesi sottosviluppati) e che la poliomielite è stata estirpata dai paesi sviluppati, riducendo comunque enormemente l'incidenza in tutto il mondo.

I no-vax compiono un atto di egoismo, contando che il (minimo) rischio lo corrono gli altri, coloro che fanno vaccinare i propri figli, creando "l'immunità di gregge".

L'immunità di gregge si crea quando in una comunità di persone o animali il numero di coloro che sono immuni è tanto elevato che, anche se la comunità viene esposta al contagio, la trasmissione dell'infezione entro la comunità si estingue in breve tempo e coinvolge pochi individui, perché la probabilità della sua trasmissione ad altri individui della comunità è trascurabile.

L'eventuale individuo colpito o guarisce o soccombe all'infezione, prima di avere il tempo e la possibilità di contaminarne un altro, vista la bassa percentuale degli individui infettabili.

Esempio 2

Molto frequenti sono le domande che riguardano v.a. che possono assumere solamente due valori.

Immaginiamo di fare una ricerca sulle preferenze di animali domestici tra uomini e donne.

	Donne	Uomini	
cani	40	60	100
gatti	60	40	100
	100	100	200

Le frequenze relative si trovano nella seguente tabella.

0.2	0.3	$\frac{1}{2}$
0.3	0.2	$\frac{1}{2}$
$\frac{1}{2}$	$\frac{1}{2}$	

Dalle frequenze marginali $\frac{N_i}{N}\frac{N^j}{N}$ ricaviamo invece la seguente tabella.

0.25	0.25	$\frac{1}{2}$
0.25	0.25	$\frac{1}{2}$
$\frac{1}{2}$	$\frac{1}{2}$	

Il procedimento per calcolare lo stimatore T è simile a quello che abbiamo già visto. Intanto, per ciascuna delle quattro caselle centrali, calcoliamo la differenza dei valori che occupano lo stesso posto.

Poi prendiamo i quadrati di queste differenze e dividiamo ciascuno dei quadrati per la frequenza relativa corrispondente (quella della seconda tabella). Sommiamo i quattro addendi e moltiplichiamo la somma per N.

I calcoli indicati ci portano al valore $T = 40$, decisamente superiore a ogni valore tabulare relativo alla $\chi^2((2-1)(2-1)) = \chi^2(1)$.

Esempio 3

Vediamo ancora un esempio di questo tipo.

Supponiamo di classificare questa volta le persone in base all'età (Giovani e Anziani) e in base alle preferenze per le vacanze: Mare o Campagna.

I dati raccolti sono riassunti nella seguente tabella.

	Giovani	Anziani	
Mare	200	100	300
Campagna	100	100	100
	300	200	500

Puoò essere conveniente anche questa volta, per rendere più trasparenti i calcoli da fare, riportare le due tabelle che abbiamo già utilizzato prima.

Le frequenze relative si trovano nella seguente tabella.

0.4	0.2	0.6
0.2	0.2	0.4
0.6	0.4	

I dati ricavabili invece dalle frequenze marginali $\frac{N_i}{N}\frac{N^j}{N}$ sono riportate nella tabella seguente.

0.36	0.24
0.24	0.16

Ne segue che

$$T = 500 \left[\frac{0.016}{0.4} + 3 \frac{0.016}{0.2} \right] = 48 \,.$$

Anche in questo caso l'indipendenza viene smentita con forza.