a cura di Francesco Romani

La legge dei grandi numeri e il Teorema del Limite Centrale

Anche se il titolo può fare paura questo articolo tratta ancora dei ritardi del Lotto. Ho notato che la maggior parte degli argomenti dei ritardisti si basano su una interpretazione errata della "Legge dei grandi numeri" e ho pensato di darne una esemplificazione con *Mathematica* citando anche il "Teorema del Limite Centrale" per chiarire ulteriormente la questione. Anche stavolta gran parte del materiale che presento è frutto delle discussioni con Dani, Elio e Adam, che, doverosamente, continuo a ringraziare.

Introduzione

Il 39 è appena uscito a Genova e la famiglia di quelli che lo avevano giocato fino all'ultimo esulta e incassa. Esultano un po' meno quelli che avevano smesso di giocare perché avevano finito i soldi oppure si erano stufati.

Cosa vuole dire che il 39 è uscito? Avevano allora ragione i ritardisti? Mettiamo subito in chiaro che i non ritardisti non dicono che il 39 (o qualunque altro numero) non uscirà mai, ma solo che giocare il 39 (o qualunque altro numero) è sempre la stessa cosa, e che giocando un numero qualunque (anche diverso tutte le settimane) si vince **in media** una volta su 18 (guadagnando però meno di 18 volte quello che si è puntato).

Cercando di studiare gli argomenti dei ritardisti più convinti mi sono imbattuto in ragionamenti basati su una concezione distorta di un complicato teorema di calcolo delle Probabilità: la cosiddetta **Legge debole dei grandi numeri** che secondo alcuni afferma che: *se un evento ha una probabilità **p** e faccio **n** esperimenti, per **n** che cresce l'evento tende a verificarsi **np** volte.* In altre parole, dicono i ritardisti: *se in media un numero deve uscire una volta su **18** ed invece tarda per **120** estrazioni allora in futuro dovrà uscire più spesso, per recuperare.*

Nel rispondere a chi fa notare che (imbrogli a parte) tutti i numeri nella cesta sono dentro sfere identiche, i ritardisti raggiungono vette di notevole poesia. Cito testualmente:

Qual è la forza che guida la mano del bambino?. La stessa che rende le bolle di sapone tonde e non cubiche. La stessa che impone che vi siano compensazioni.

Se un numero ritarda molto, nel grafico si formerà un pozzo, e se la teoria è esatta il pozzo deve essere riempito al più presto possibile per avere sempre una linea orizzontale più dritta possibile. Altrimenti se il pozzo persiste nel tempo significherebbe che quel numero ha meno probabilità degli altri di uscire (quindi la forza c'è).

In questo articolo voglio provare a far vedere cosa dice **davvero** la legge dei grandi numeri e approfondisco la questione andando a pescare anche il **Teorema del Limite Centrale**. Il prossimo paragrafo è un po' duro da leggere per chi non gradisce le radici quadrate e gli integrali, ma poi cercheremo di verificare sul campo con *Mathematica* se la pratica corrisponde alla teoria.

Un po' di Probabilità

Consideriamo una variabile casuale **X** con **k** possibili realizzazioni $x_1, x_2, \ldots, x_k$, e probabilità associate $p(x_1), p(x_2), \ldots, p(x_k)$. La media o valore aspettato di **X** è definita come

$$\mu = E(X) = \sum_{i=1}^{k} x_i p(x_i).$$

La varianza è definita come

$$\sigma^2 = E(X-\mu)^2 = \sum_{i=1}^{k} (x_i - \mu)^2 p(x_i).$$

Esempio: se giochiamo a Testa o Croce con una moneta "onesta", possiamo associare a **Testa** il valore **1** e a **Croce** il valore **-1**, allora **k = 2**, $x_1 = 1$, **x2 = -1**, $p(x_1) = p(x_2) = 1/2$. La media è $\mu = 1/2 - 1/2 = 0$ e la varianza $\sigma^2 = 1/2 + 1/2 = 1$.

Se assegnassimo a **Testa** e **Croce** valori numerici diversi (per esempio **0** e **1**) avremmo conti più complicati ma alla fine i risultati sarebbero perfettamente equivalenti.

Per definizione due variabili casuali sono **indipendenti** se in nessun modo la conoscenza del risultato di una di esse può influenzare le previsioni del risultato dell'altra.

Siano $X_1, X_2, \ldots, X_n, \ldots$ variabili casuali **indipendenti** tutte con la stessa distribuzione di probabilità e siano μ e σ^2 rispettivamente la loro media e la varianza. Esistono due importanti teoremi, di non facile dimostrazione, che studiano il comportamento della quantità $S_n = X_1+X_2+\ldots+X_n$.

La legge debole dei Grandi Numeri (LDGN)

Per ogni $\varepsilon > 0$ prefissato piccolo a piacere, vale

$$\lim_{n \to \infty} \Pr.\left(\left| \frac{S_n}{n} - \mu \right| > \varepsilon \right) = 0$$

Cosa dice questo Teorema? La quantità S_n/n per valori grandi di **n** tende a μ con probabilità **1**!

Nel nostro esempio se **T** è il numero delle teste e **C** il numero delle croci **LDGN** dice che $S_n/n = (T-C)/n$ tende a zero con **n**.

Si noti che **LDGN** non dice nulla sul comportamento di $S_n = n(T-C)$. Se quest'ultima quantità andasse a zero avrebbero ragione i ritardisti ma in realtà S_n può tranquillamente andare all'infinito mentre S_n/n va a zero!! Per vedere come si comporta S_n bisogna scomodare un teorema dall'enunciato molto più indigesto.

Il Teorema del Limite Centrale (CLT)

Per ogni $\varepsilon > 0$ prefissato piccolo a piacere, vale

$$\lim_{n \to \infty} \Pr.\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt, \quad -\infty < x < \infty$$

Il Teorema scritto in questo modo è un risultato fondamentale in Teoria della Probabilità ma dice pochino ai fini della polemica sui ritardi. Con un po' di manipolazioni elementari si ottiene però:

$$\lim_{n \to \infty} \Pr.\left(|S_n - n\mu| \le x\sigma\sqrt{n} \right) = \operatorname{erf}\left(\frac{x}{\sqrt{2}} \right), \quad 0 \le x < \infty$$

Dove **erf(x)** è la curva degli errori definita come

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$$

e (guarda caso) già implementata in *Mathematica* come **`Erf[x]`**. Vale **erf(0)=0** e **erf(x)** tende rapidamente ad **1** per **x** che va all'infinito, come si intuisce dalla definizione e si vedrà nell'ultima figura.

Cosa dice **CLT** in questa forma? Essenzialmente che $|S_n - n\mu|$ non cresce più velocemente della radice quadrata di **n** moltiplicata per la varianza della distribuzione e per una costante **x** che tiene conto della probabilità con cui vogliamo che la limitazione sia valida.

Nel caso del gioco Testa o Croce, sostituendo la media **0** e la varianza **1** si ha

$$\lim_{n \to \infty} \Pr.\left(|S_n| \le x\sqrt{n} \right) = \operatorname{erf}\left(\frac{x}{\sqrt{2}} \right)$$

che per **x=2** diviene

$$\lim_{n \to \infty} \Pr.\left(|S_n| \le 2\sqrt{n} \right) = 0.9545\ldots$$

mentre per **x=3** diviene

$$\lim_{n \to \infty} \Pr.\left(|S_n| \le 3\sqrt{n} \right) = 0.9973\ldots$$

In altre parole tutto quello che si può dire su S_n è che non va all'infinito più velocemente di una costante per la radice di **n**. Ovviamente dividendo per **n** si ottiene come corollario la **LDGN.** Si tenga presente che ragionando in termini assoluti e non probabilistici l'unica limitazione che si può dare è $|S_n| \le n$., ovvero potrebbero anche uscire tutte teste o tutte croci, (un calcolo diretto, il buonsenso oppure anche il **CLT** ci assicurano che questo evento è molto raro).

Ora basta con la teoria e vediamo di fare qualche simulazione.

Giochiamo a Testa e Croce

Scriviamo una funzione che lancia le monete e tiene conto delle teste e delle croci: la variabile **`nt`** contiene la quantità **T-C**. Teniamo conto del numero **x** di lanci effettuati e disegniamo il grafico di **`nt`** al variare di **`x`**. Il programma seguente fa **`n`** gruppi di **`m`** lanci e disegna un grafico nel colore **`col`**.

In[1]:=

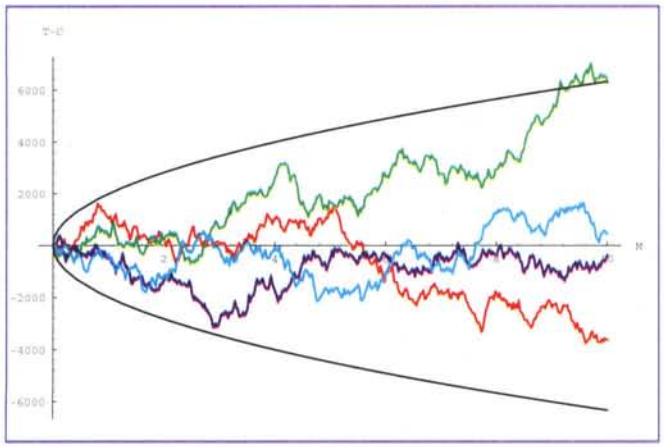
```
prova[col_]:=(
 nt=0;
 x=0;
 LL={};
 Do[nt+=2(Plus@@Table[Random[Integer,{0,1}],{m}])-
m;
   x+=m;
   AppendTo[LL,{x 10^-6,nt}],{n}];
 ListPlot[LL,
     PlotStyle->col,
     PlotJoined->True]);
```

Facciamo 10 milioni di lanci per quattro volte con quattro colori diversi e disegniamo anche il grafico di $\pm 2\sqrt{x}$ ovvero l'area che in base al **CLT** ha una probabilità **erf(2) = 0.95...** di essere riempita.

In[2]:=

```
n=400;
m=25000;
p1=Plot[ 2Sqrt[x 10^6],
   {x,0,m n 10^-6}];
p2=Plot[-2Sqrt[x 10^6],
   {x,0,m n 10^-6}];
lR=prova[Red];
lG=prova[Green];
lB=prova[Blue];
lC=prova[Cyan];
Show[lR,lG,lB,lC,p1,p2,
   PlotRange->All,
   AxesLabel->{"M","T-C"}]
```

Vedi qui sotto la **Figura 1**



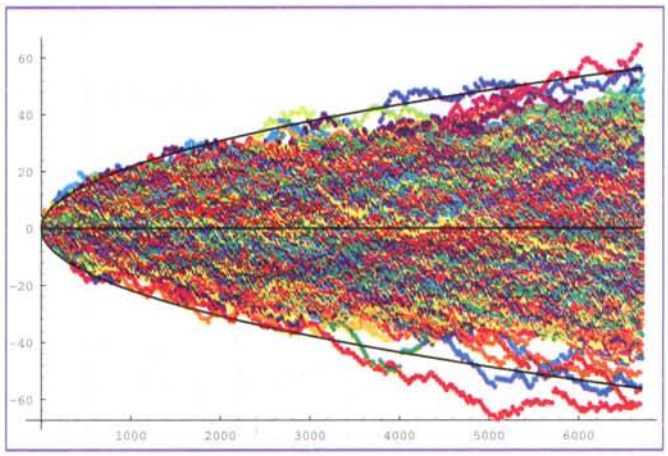
Talvolta il numero delle teste uguaglia il numero delle croci ma altre volte la curva si allontana allegramente dall'asse delle ascisse. Tenete presente che poiché la radice di **x** cresce più lentamente di **x** in tutti e **4** i casi la media va a zero.

Il Lotto

Vediamo ora cosa succede con il lotto, abbiamo **90** numeri ognuno dei quali ha probabilità **1/18** di essere estratto. In altre parole se **e(i)** rappresenta il numero di volte che viene estratto **i**, **LDGN** assicura che **e(i)/n** tende a **1/18**, ma nessuno può affermare che **e(i)** tenda a **18n**. La quantità **e(i)-18n** ci dice quante volte in più o in meno il numero **i** è uscito rispetto al valore atteso **18n**. Tutto quello che ci assicura **CLT**, invece è che **e(i)-18n** non cresce tipicamente più di una costante per la radice di **n**.

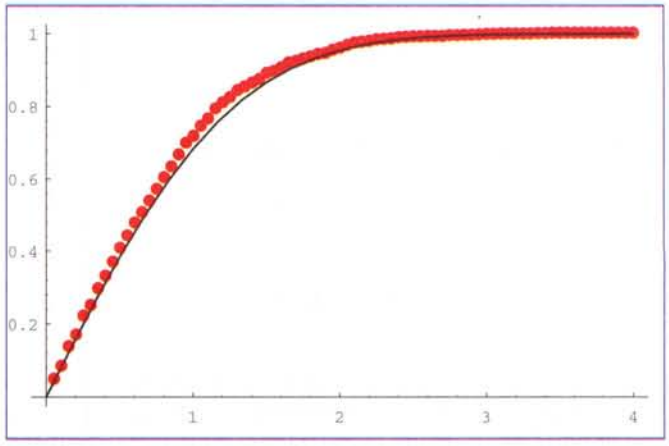
Stavolta invece di fare una simulazione ho preso i file delle estrazioni del lotto storiche (circa 6000 estrazioni per più di cento anni di gioco). Considerando indipendenti le uscite sulle 10 ruote abbiamo 900 numeri distinti ognuno dei quali percorre una traiettoria di un colore diverso.

Nella **Figura 2** riportata qui sotto il numero di uscite in ritardo o in anticipo sulla media di ogni numero è disegnato in funzione del numero di estrazioni. La parabola in nero racchiude l'area che in base al **CLT** ha una probabilità **erf(3) = 0.997...** di essere riempita.



È straordinariamente evidente come lo spazio permesso dalla teoria venga quasi tutto riempito. Vi sono numeri che dopo **6000** estrazioni sono in ritardo o in anticipo di **60** uscite sulla media. Per **n** che cresce alcuni di questi divari vanno a zero ma altri aumentano, (per la cronaca ho fatto una simulazione di una ruota per **2 milioni** di estrazioni ottenendo un grafico con lo stesso aspetto).

È anche possibile verificare direttamente il **CLT** confrontando il numero di ritardi o anticipi di un certo valore misurati dopo tutte le oltre 6000 estrazioni con quelli previsti. La **Figura 3** confronta il grafico di **erf(x)** (in nero) con quanto ottenuto storicamente per tutti i **900** numeri (pallini rossi). L'accordo è molto buono! MC



romouse

22
30
maggio '99

Fiera di Roma

Romouse punta all'affermazione della Capitale quale valido scenario di eventi ad ampio raggio in area informatica.

Tecnologie evolute, novità hardware e software per l'utenza privata e quella aziendale, multimedia, CAD, EDP, internet e telecomunicazioni per questo salone che nasce quale riferimento nuovo ed evoluto in risposta all'esigenza del centro-sud di trovare in una sola manifestazione tutte le novità del mercato informatico ed una piattaforma credibile di analisi verso cui possano convergere le aziende e le professionalità più rappresentative.

romouse

In concomitanza con:

MOA CASE&COSE

www.moacasa.com
info@moacasa.com

organizzazione tecnica:
Parisse Pubblicità
Tel. 0630891701
Fax 0630892034
E-mail: pparisse@iol.it