LA LEGGE DELLE PAGINE GUALCITE

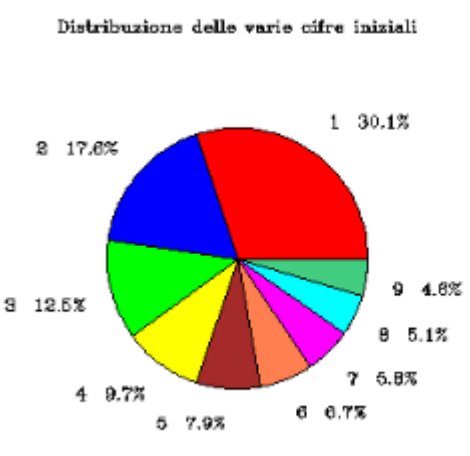
Saper fare a mente le moltiplicazioni di due numeri che incominciano per 1, caso particolarmente semplice, non è del tutto futile, ed è assai più utile che saper fare a mente moltiplicazioni di due numeri che incominciano per 9.

A occhio e croce diremmo che i numeri hanno la stessa probabilità di incominciare per 1 o per 2 o per 3 o per 4 o per 9. Di certo, numeri presi a caso funzionano così. Però, se i numeri si riferiscono a una classe di oggetti, per esempio lunghezze di fiumi, altezze di montagne, lunghezze di poemi in versi, valore di azioni in borsa, o parole di un poema eccetera, si scopre che lunghezze, altezze etc. che incominciano per 1 sono più frequenti, circa il 30%, invece che 1/9 (11%) del totale.

Questo fatto abbastanza controintuitivo fu notato da varie persone che utilizzavano libri di tavole numeriche (oggi non si usano quasi più) e osservavano che le prime pagine erano più gualcite - più usate - delle ultime. F. Benford pensò bene di raccogliere i dati rilevanti nel 1938, e pubblicò la sua "legge" in quell'anno. Era stato preceduto da S. Newcomb, nel 1881, ma assai probabilmente non lo sapeva.
Naturalmente c'è anche una spiegazione matematica, anzi, ce ne sono troppe, perché la legge non si applica sempre allo stesso modo, in quanto dipende dalla distribuzione dei dati, che, idealmente, dovrebbe essere uniforme. Noi ne daremo in appendice una dimostrazione basata sull'invarianza di scala, cioè, in pratica, sul fatto che la distribuzione, delle prime cifre, ad esempio nelle lunghezze dei fiumi, altezze di montagne, non deve cambiare in forma se noi misuriamo le lunghezze in metri, miglia, o chilometri o leghe.

Darò ora un'indicazione intuitiva di perché la legge funziona.
La tabella delle percentuali dei numeri che si riferiscono ad una data classe di oggetti e che incominciano con una data cifra (omettendo lo zero, che ci rimanda alla prima cifra significativa) è la seguente:

Prima cifra	Percentuale
1	30%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.6%
8	5.1%
9	4.6%



Trovare la legge esatta non è immediato. Ma possiamo renderci conto di quello che succede. Supponiamo di avere una classe di oggetti (come ad esempio i fiumi di un dato Paese) a ciascuno dei quali è assegnato un numero (come ad esempio la lunghezza dei fiumi). Mentre la legge delle pagine gualcite riguarda la prima cifra, la chiave per capire come ciò avvenga è concentrarsi sui valori massimi della classe di numeri che ci interessa.

Se, in un gruppo di fiumi scelti a caso, la lunghezza minima di un fiume è 1 Km e la massima è 2000 Km, i numeri tra questi 2000 che incominciano per 1 sono 1111 (Quali?). Quindi scegliendo a caso un numero fra 1 e 2000 abbiamo più di metà della probabilità che il numero incominci per 1. Se il massimo è 3000, la probabilità che incominci per 1 è comunque ancora 30% . Se il massimo è 4000 scendiamo a 25%. Ma non potremo scendere sotto 11%, che raggiungeremo solo quando il massimo sarà 10000.

Supponiamo invece di cercare i numeri che incominciano per 9. Per un massimo che vale 1000, i numeri che incominciano per 9 sono 111, che restano tali, con probabilità quindi decrescenti (111/2000, poi 111/3000, poi 111/4000 etc.) , fino a che non arriviamo ad un numero finale 10000, nel qual caso ne avremo in media 1111 come per tutti gli altri numeri.
Naturalmente se il massimo numero è 700 000, vale lo stesso ragionamento. C'è qualche zero in più, ma le proporzioni di numeri che incominciano per una data cifra sono le stesse.
E' chiaro d'altra parte che la distribuzione delle seconde cifre, pur non essendo uniforme (uguale probabilità per tutte le cifre), lo è di più di quella delle prime cifre, e diventa sempre più uniforme nelle cifre successive.

Bisogna osservare che, per quanto la legge di Benford sia derivabile matematicamente (almeno fino ad un certo punto), ci sono dei casi pratici in cui essa non è rispettata. Quindi non bisogna stupirsi se la distribuzione osservata non si accorda sempre con la tabella data sopra. Il caso ideale in cui l'accordo è quasi perfetto è quello di una distribuzione di numeri (1) abbastanza uniforme (2) su diversi ordini di grandezza. Per esempio, la distribuzione della statura degli uomini adulti non può seguire appieno la legge di Benford, poiché tende a essere limitata fra due estremi abbastanza vicini.

Non si creda troppo alla derivazione "analitica" che daremo più sotto. Ad ogni modo, che una "distribuzione delle distribuzioni" segua la legge di Benford, fu dimostrato rigorosamente da Hill nel 1998. Questo spiega in certo senso perché, nonostante molte distribuzioni non abbiano un andamento logaritmico, tuttavia le pagine delle tavole numeriche restino logaritmicamente gualcite (ad ammuffire malinconicamente, perché, dopo tanta fatica per calcolarle, chi le usa più?)

Tuttavia, pur con queste restrizioni, la legge ha molte applicazioni, una volta che si dimostri che è valida nel caso in esame. Ad esempio frodi fiscali, elettorali, scientifiche (alterazione dei dati) possono essere messe in luce con questa legge, che sembra si applichi soprattutto alle frodi.

APPENDICE NUMERICA

A scopo dimostrativo, si può provare a usare il programma SmallBasic che accludo (se si scarica il testo in .pdf e poi lo si converte in formato .txt o anche .docx, lo si può poi copiare direttamente sul programma SmallBasic che sarà stato scaricato seguendo le istruzioni date altrove in questo

sito: http://dainoequinoziale.it/scienze/matematica/2016/12/16/usodismallbasic.html).

```
'Pagine gualcite
TextWindow.WriteLine("Legge di Benford o delle pagine gualcite")
start:
TextWindow.WriteLine("Numero massimo?")
massimo =TextWindow.ReadNumber()
TextWindow.WriteLine("Quanti tentativi?")
count1=0
count2=0
count1=0
count3=0
count4=0
count5=0
count6=0
count7=0
count8=0
count9=0
fine= TextWindow.ReadNUmber()
For I = 1 To Fine
  nn = Math.GetRandomNumber (massimo)
  If Text.StartsWith(nn, 1) Then
    count1 = count1+1
  EndIf
  If Text.StartsWith(nn, 2) Then
    count2 = count2+1
  EndIf
  If Text.StartsWith(nn, 3) Then
    count3 = count3+1
  EndIf
  If Text.StartsWith(nn, 4) Then
    count4 = count4+1
  EndIf
  If Text.StartsWith(nn, 5) Then
    count5 = count5+1
  EndIf
  If Text.StartsWith(nn, 6) Then
    count6 = count6+1
  EndIf
  If Text.StartsWith(nn, 7) Then
    count7 = count7+1
  EndIf
  If Text.StartsWith(nn, 8) Then
    count8= count8+1
  EndIf
  If Text.StartsWith(nn, 9) Then
    count9 = count9+1
  EndIf

  EndFor
```

```
  TextWindow.WriteLine("Con massimo" + massimo + " su " + fine + "
tentativi, i numeri che incominciano con 1 sono " + count1)
  TextWindow.WriteLine("Con massimo " + massimo + " su " + fine + "
tentativi, i numeri che incominciano con 2 sono " + count2)
  TextWindow.WriteLine("Con massimo " + massimo + " su " + fine + "
tentativi, i numeri che incominciano con 3 sono " + count3)
  TextWindow.WriteLine("Con massimo " + massimo + " su " + fine + "
tentativi, i numeri che incominciano con 4 sono " + count4)
  TextWindow.WriteLine("Con massimo " + massimo + " su " + fine + "
tentativi, i numeri che incominciano con 5 sono " + count5)
  TextWindow.WriteLine("Con massimo " + massimo + " su " + fine + "
tentativi, i numeri che incominciano con 6 sono " + count6)
  TextWindow.WriteLine("Con massimo " + massimo + " su " + fine + "
tentativi, i numeri che incominciano con 7 sono " + count7)
  TextWindow.WriteLine("Con massimo " + massimo + " su " + fine + "
tentativi, i numeri che incominciano con 8 sono " + count8)
  TextWindow.WriteLine("Con massimo " + massimo + " su " + fine + "
tentativi, i numeri che incominciano con 9 sono " + count9)
Goto start
```

Il programma vi lascia scegliere il numero massimo (per esempio la lunghezza del fiume più lungo, in numero di abitanti della città più popolata etc.), a cui darà il nome "nn"
Inoltre vi lascerà scegliere il numero di "tentativi"(a cui darà il nome "fine").
A caso sceglierà un numero "fine " di numeri e conterà quelli tra loro che incominciano con una data cifra minore di nn.
Potete anche modificare il programma in modo da lasciare che sia il computer a scegliere a caso il numero massimo nn, per esempio mettendo nella quinta riga:

```
massimo = Math.GetRandomNumber (1000)
```

Se fisserete ad esempio il massimo, nn, a 700, vedrete che i numeri scelti a caso si divideranno più o meno egualmente tra 1 e 6, ma lasceranno assai poco per 7, 8, 9. E se mi sarò spiegato bene, la cosa non vi sorprenderà.

Il lettore che si intende di programmi noterà che qui il numero minimo è 0. Però potrà modificare il programma leggermente in modo da partire da un numero minimo, per esempio la lunghezza minima di un fiume tra quelli considerati.

APPENDICE ANALITICA per chi conosce il calcolo integrale

Una dimostrazione (almeno illusoriamente) più rigorosa è data nel seguente modo: la distribuzione di probabilità P(n) che un numero incominci con una data cifra n (ripeto, non un numero a caso, ma la cifra iniziale, per esempio, dell'altezza di una montagna facente parte di un gruppo) dovrebbe essere indipendente da un fattore di scala. Per esempio, possiamo misurare la lunghezza di un fiume in pollici, kilometri o miglia, l'altezza delle montagne in yarde, metri, etc. Ma la distribuzione generale dovrebbe essere invariante.

Passando al continuo, l'invarianza di scala richiede che la distribuzione delle prime cifre P(x) non dipenda dal fattore k di scala. Quindi, grossolanamente deve essere:

$$P(kx) = f(k)\ P(x)$$

Per esempio, se in una yarda ci sono 3 piedi, ci aspettiamo che la probabilità che la prima cifra sia 1 in yarde, sia eguale alla probabilità che una misura in piedi dia 3,4,5 (rispettivamente 1.33, 1.66, 1.99).

L'integrale della distribuzione $\int P(x)\,dx = 1 = \int P(kx)d(kx)$ richiede, moltiplicando a numeratore e denominatore per k, che $\int P(kx)dx = 1/k$, il che comporta $f(k) = 1/k$. Quindi

$$P(kx) = \frac{1}{k}\,P(x)$$

Ora si differenzia la relazione sopra riportata rispetto a k, e si trova:

$$x\,P'(kx) = -\left(\frac{1}{k^2}\right)P(x)$$

che deve valere anche per k = 1. Cioè:

$$\frac{P\prime(x)}{P(x)} = -\frac{1}{x}\text{ , cioè }\quad \frac{dy}{y} = -\frac{dx}{x}\quad \text{da cui } \ln y = -\ln x + \ln C \quad \text{e } y = \frac{1}{x}$$

Ci sarebbe in verità una costante moltiplicativa, ma noi la includeremo nel fattore di scala. Come probabilità di distribuzione questa lascia un poco a desiderare, perché l'integrale diverge. Ma nei dati che ci troviamo a maneggiare, normalmente ci sono dei limiti superiori e inferiori: non ci sono monti più alti dell'Everest o fiumi più lunghi del Nilo.
Se si integra tra due cifre n, n+1, si ottiene la percentuale

$$P_n = \frac{\int_n^{n+1} P(x)\,dx}{\int_1^{10} P(x)\,dx} = \log_{10}\left(1+\frac{1}{n}\right)$$

Come si è detto, non si creda troppo a questa derivazione.