

Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#432 Scale invariance (power law, Zipf, Benford, Pareto, etc.)

été 2020

Scale Invariance (and Fractals)



Heuristics

Scale invariance means that X and λX should have the same distribution,

$$f(\lambda x) \propto f(x), \quad \forall \lambda > 0 \text{ and } x \geq 1.$$

The only possibility is $f(x) \propto x^{-\gamma}$ (power law).

Note: finite expected value if $\gamma > 2$, finite variance if $\gamma > 3$

Alternatively, it could mean that (if $\bar{F} = 1 - F$)

$$\bar{F}(\lambda x) \propto \bar{F}(x), \quad \forall \lambda > 0 \text{ and } x \geq 1.$$

Then $\bar{F}(x) \propto x^{-\alpha}$ (power law), see [wikipedia](#)

(see also memoryless property for another scale invariance, #433)

Pareto distribution

60

VILFREDO PARETO

Vilfredo Pareto *La Legge della Domanda, Gionale degli Economisti*, 1895 or *La courbe de la répartition de la richesse, Université de Lausanne*, 1896.

$$\mathbb{P}[X > x] = x^{-\alpha}, \quad x \geq 1.$$

or

$$\log \mathbb{P}[X > x] = -\alpha \log(x), \quad x \geq 1.$$

with here $\alpha = \gamma - 1$.

Si dà la combinazione che si può ottenere una legge empirica assai semplice per quella distribuzione delle entrate. Per non dilungarci troppo non daremo qui tutti i particolari del calcolo che saranno tra breve pubblicati nel nostro Corso di Economia Politica, e ci limiteremo a un breve cenno.

Sia x l'entrata di ciascun capo di famiglia, o meglio di ciascun contribuente

y dx il numero di capi di famiglia, o di contribuenti aventi un'entrata compresa tra x e $x + dx$. Si osserva che i numeri dati da molte statistiche possono figurarsi con la formola

$$(1) \quad y = \frac{H}{x^\alpha}$$

La Sassonia è uno dei paesi per quali si hanno migliori statistiche delle entrate dei cittadini. Lo specchio seguente ci darà un'idea della approssimazione della formola (1)

L'équation de cette ligne peut se représenter par

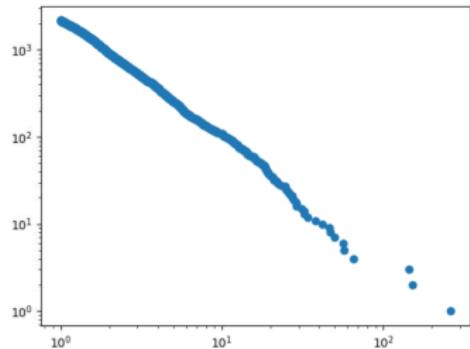
$$(1) \quad \text{Log } N = \text{Log } A - \alpha \text{ Log } x : \\ \text{ce qui donne}$$

$$(2) \quad N = \frac{A}{x^\alpha}.$$

Pareto Plot

```
1 > import pandas as pd  
2 > import matplotlib.pyplot  
     as plt  
3 > url="http://  
      freakonometrics.free.fr/  
      danish.csv"  
4 > base=pd.read_csv(url,sep  
     =";",decimal=",")  
5 > x = base.loss  
6 > x = x.sort_values()  
7 > n = len(x)  
8 > y = range(1,n+1)  
9 > y = list(reversed(y))  
10 > plt.plot(x,y,'o')  
11 > plt.xscale('log')  
12 > plt.yscale('log')  
13 > plt.style.use('seaborn-  
     whitegrid')  
14 > plt.show()
```

```
1 > url="http://  
      freakonometrics.free.fr/  
      danish.csv"  
2 > base=read.csv(url,sep  
     =";",dec=",")  
3 > x=sort(base$loss)  
4 > n=length(x)  
5 > y=(n:1)  
6 > plot(x,y,log="xy")
```

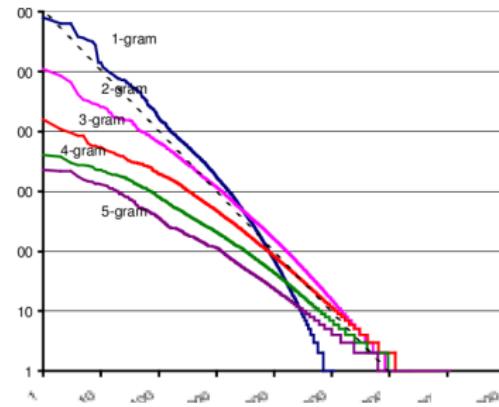
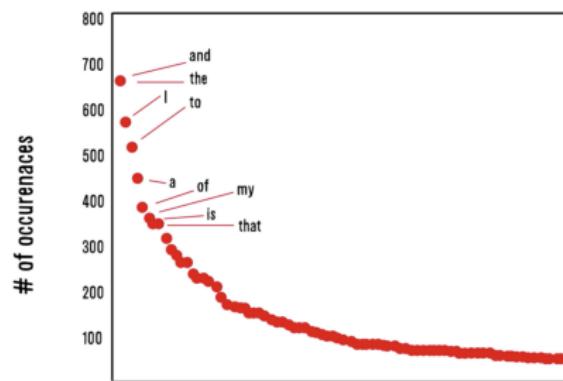


Zipf's Law

$$p(x) \propto \frac{1}{x^\gamma}, \quad x = 1, 2, \dots, n \quad (\text{for some integer } n)$$

see [Statistical laws in linguistics](#) by Vitold Belevitch.

Use on word counts (via Zipf's Law)



but also k -grams (see WSJ, [Extension of Zipf's Law to Words and Phrases](#))

Scale-Free Network

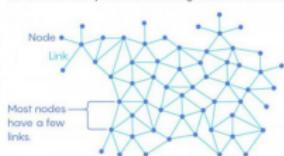
To Be or Not to Be Scale-Free

Scientists study complex networks by looking at the distribution of the number of links (or "degree") of each node.

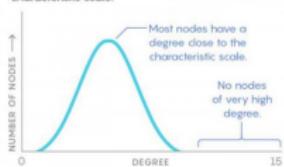
Some experts see so-called scale-free networks everywhere. But a new study suggests greater diversity in real-world networks.

Random Network

Randomly connected networks have nodes with similar degrees. There are no (or virtually no) "hubs"—nodes with many times the average number of links.

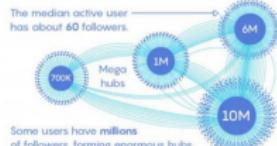


The distribution of degrees is shaped roughly like a bell curve that peaks at the network's "characteristic scale."

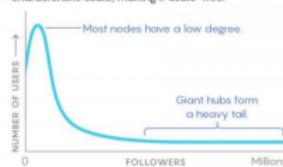


Twitter's Scale-Free Network

Most real-world networks of interest are not random. Some nonrandom networks have massive hubs with vastly higher degrees than other nodes.

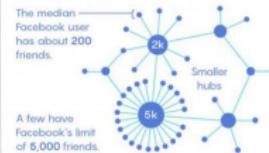


The degrees roughly follow a power law distribution that has a "heavy tail." The distribution has no characteristic scale, making it scale-free.

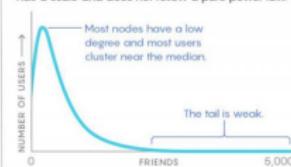


Facebook's In-Between Network

Researchers have found that most nonrandom networks are not strictly scale-free. Many have a weak heavy tail and a rough characteristic scale.



This network has fewer and smaller hubs than in a scale-free network. The distribution of nodes has a scale and does not follow a pure power law.

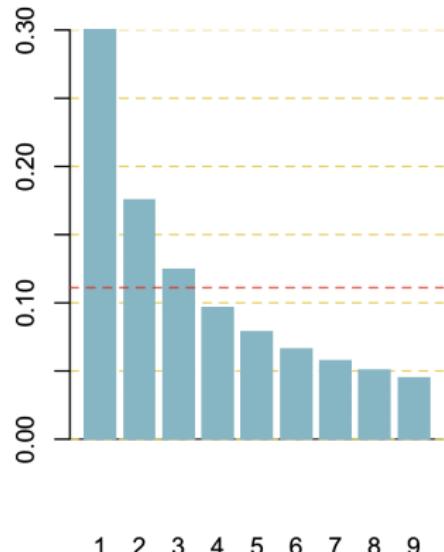
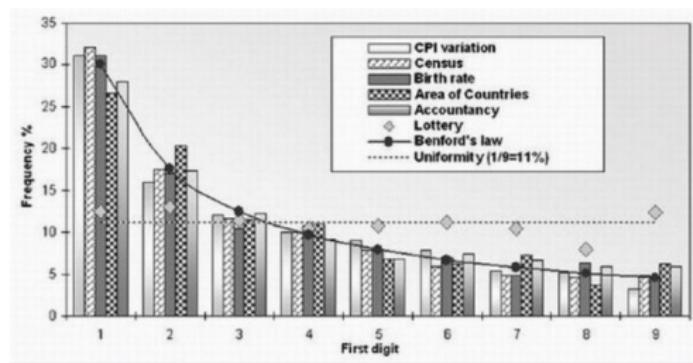


Source: <https://colorado.edu/biofrontiers/>, from Broido & Clauset (2018)

(Newcomb-) Benford's Law

From [The law of anomalous numbers](#), Benford (1938)

$$\begin{aligned} p(d) &= \log_{10}(x+1) - \log_{10}(x) \\ &= \log_{10}\left(1 + \frac{1}{x}\right) \\ \text{with } x &= 1, 2, \dots, 9 \end{aligned}$$



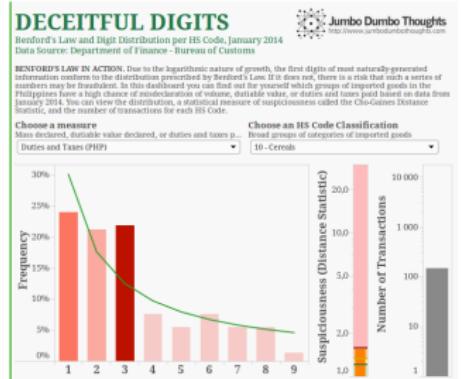
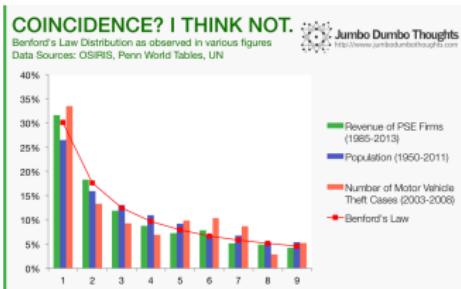
via [What is Benford's Law and why is it important for data science?](#), see also [COVID-19](#)

(Newcomb-) Benford's Law

Not *per se* a power law, but appears when the underlying variable has a power distribution, see *L'étonnante loi de Benford*

Use in many application, especially in fraud detection, see *Benford's Law and the Detection of Election Fraud*, Deckert, Myagkov & Ordeshook (2011)

Note : can be extended to more digits



Source: On Benford's Law: Determining import fraud risk using customs data