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,

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

The only possiblity 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 $\overline{F} = 1 - F$)

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

Then $\overline{F}(x) \propto x^{-\alpha}$ (power law), see wikipedia
(see also memoryless property for another scale invariance, #433)

Pareto distribution

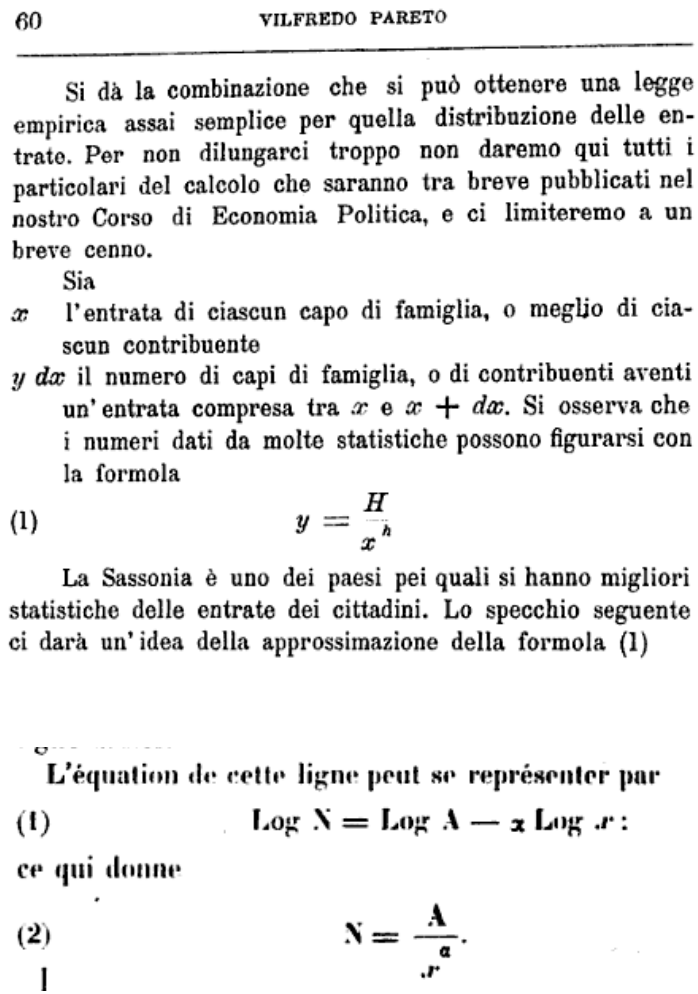
Vilfredo Pareto La Legge della Domanda, *Gionale degli Economisti*, 1895 or La courbe de la répartition de la richesse, *Université de Lausanne*, 1896.

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

or

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

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



60 VILFREDO PARETO

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) $$y = \frac{H}{x^h}$$

La Sassonia è uno dei paesi pei 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) $$\text{Log } N = \text{Log } A - \alpha \text{ Log } x :$$

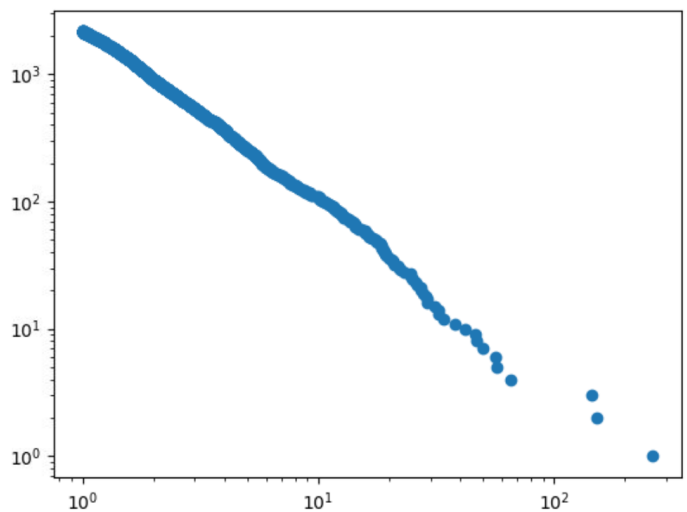
ce qui donne

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

Pareto Plot

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

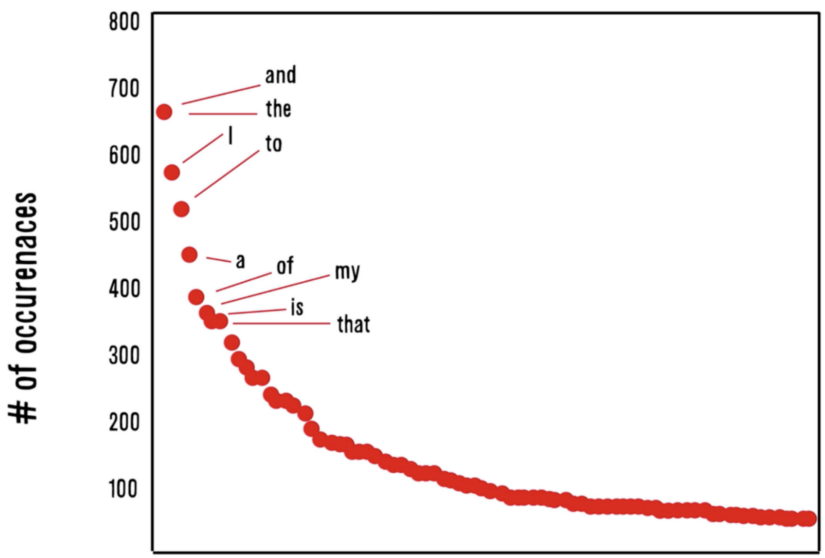
```
> url="http://
    freakonometrics.free.fr/
    danish.csv"
> base=read.csv(url,sep
    =";",dec=",")
> x=sort(base$loss)
> n=length(x)
> y=(n:1)
> plt(x,y,log="xy")
```

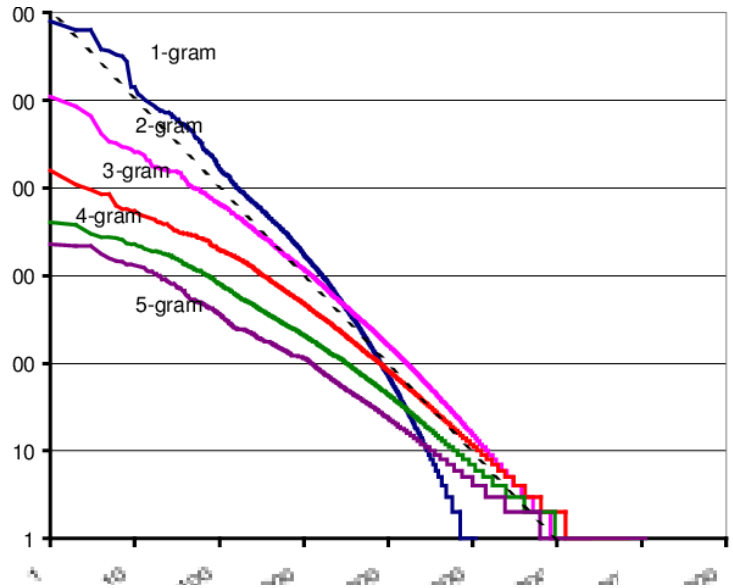


Zipf's Law

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

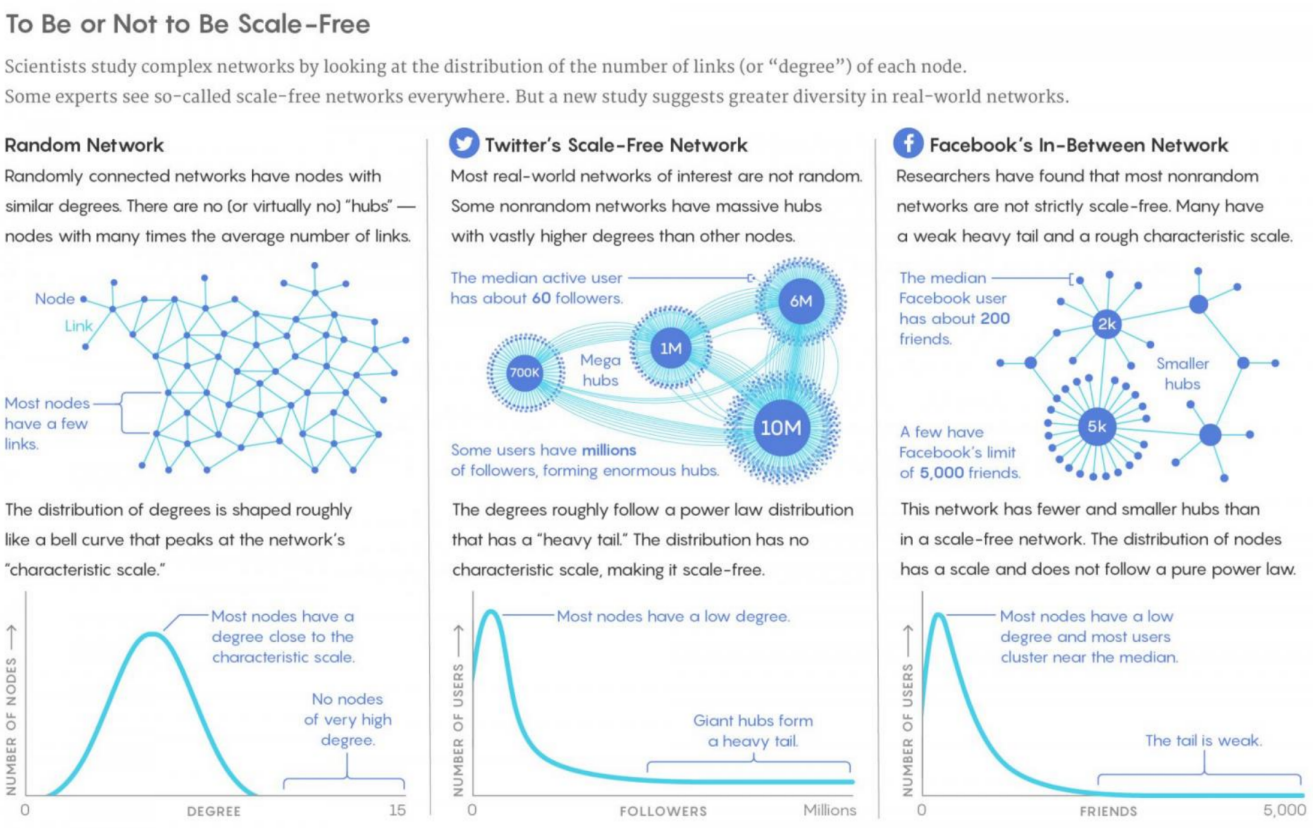
see Statistical laws in linguistics by Vitold Belevitch.
Use on word counts (via Zipf's Law





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

Scale-Free Network



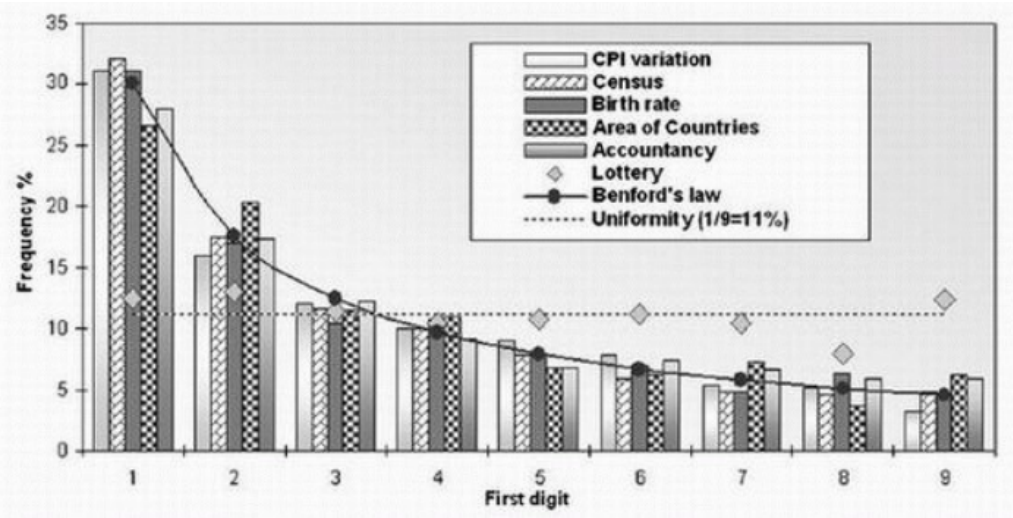
Source: https://colorado.edu/biofrontiers/, from Broido & Clauset (2018)

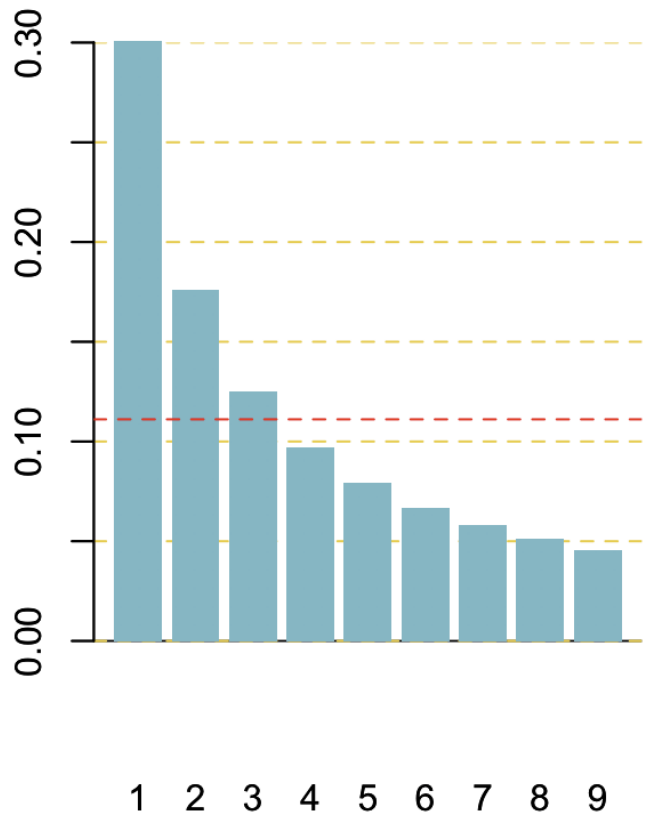
(Newcomb-) Benford's Law

From The law of anomalous numbers, Benford (1938)

$$p(d) = \log_{10}(x+1) - \log_{10}(x)$$
$$= \log_{10}\left(1 + \frac{1}{x}\right)$$

with $x = 1, 2, \cdots, 9$





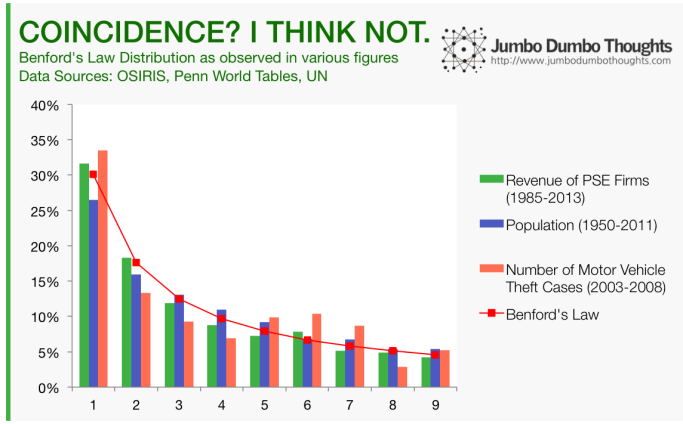
via What is Benford's Law and why is it important for data science?, see also COVID-19

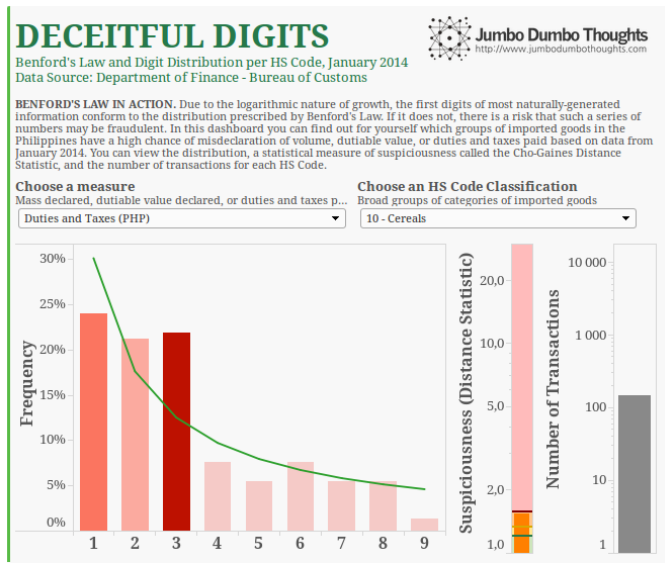
(Newcomb-) Benford's Law

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