Statistical Methods for Data Science

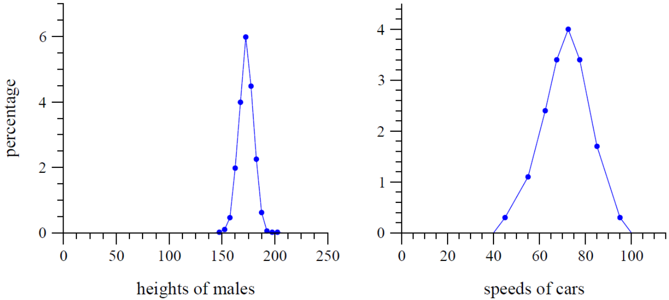
Lesson 08 - Power laws and Zipf laws.

Salvatore Ruggieri

Department of Computer Science
University of Pisa
salvatore.ruggieri@unipi.it

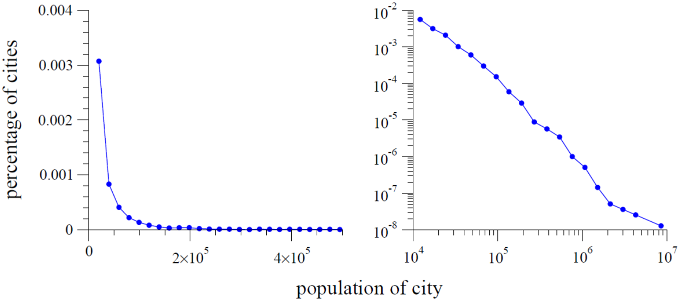
Scaled distributions

- Many of the things that scientists measure have a typical size or “scale” — a typical value around which individual measurements are centered



Scale-free distributions

- But not all things we measure are peaked around a typical value. Some vary over an enormous dynamic range.



Look at **Figure 4 of Newman's paper**

Continuous power-law

Power-law

A continuous random variable X has the *power-law distribution*, if for some $\alpha > 1$ its density function is given by

$$p(x) = C \cdot x^{-\alpha} \quad \text{for } x \geq x_{min}$$

We denote this distribution by $Pow(x_{min}, \alpha)$.

- C is called the **intercept**, and α the **exponent**.
- Passing to the logs:

$$\log p(x) = -\alpha \cdot \log(x) + \log C$$

linearity in log-log scale plots!

See R script

Intercept

- What is the constant C?

$$1 = \int_{x_{min}}^{\infty} Cx^{-\alpha} dx = \frac{C}{1-\alpha} \left[x^{-\alpha+1}\right]_{x_{min}}^{\infty} = \frac{C}{1-\alpha} \left[\infty^{-\alpha+1} - x_{min}^{-\alpha+1}\right] = \frac{C}{\alpha-1} x_{min}^{-\alpha+1}$$

- Finite only for $\alpha > 1$, and then:

$$C = (\alpha - 1) x_{min}^{\alpha-1}$$

- In summary:

$$p(x) = \frac{(\alpha-1)}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\alpha}$$

CCDF

- Let's compute:

$$P(X > x) = \int_x^{\infty} p(y)dy = C\int_x^{\infty} y^{-\alpha}dy = \frac{C}{1-\alpha}\left[y^{-\alpha+1}\right]_x^{\infty} = \frac{C}{\alpha-1}x^{-\alpha+1}$$

- and since $C = (\alpha - 1)x_{min}^{\alpha-1}$:

$$P(X > x) = \left(\frac{x}{x_{min}}\right)^{-\alpha+1}$$

- Same form as the PDF with exponent $(\alpha - 1)$ and no normalization constant!

See R script

Scale-free distributions

$$p(bx) = g(b)p(x)$$

- Measuring in cm, inches, Km, or miles does not change the form of the distribution (up to some constant)!
- For a power-law $p(x) = Cx^{-\alpha}$

  $$p(bx) = b^{-\alpha} Cx^{-\alpha}$$

  hence, $g(b) = b^{-\alpha}$
- Actually, power-laws are the only scale-free distributions!
  - see Eq. 30-34 of **Newman's paper** for a proof

Pareto distribution

Pareto

A continuous random variable X has the *Pareto distribution*, if for some $\beta > 0$ its density function is given by

$$p(x) = C \cdot x^{-(\beta+1)} \quad \text{for } x \geq x_{min}$$

We denote this distribution by $Par(x_{min}, \beta)$.

- $Par(x_{min}, \beta) = Pow(x_{min}, \beta + 1)$ or $Pow(x_{min}, \alpha) = Par(x_{min}, \alpha - 1)$
- Pareto noticed that the number of people whose income exceeded level x (i.e., CCDF) was well approximated by C/x^{β} for some constants C and $\beta > 0$
- It appears that for all countries $\beta \approx 1.5$.
- In formula, CCDF of $Par(x_{min}, \beta)$ is $(\frac{x}{x_{min}})^{-\beta-1+1} = (\frac{x}{x_{min}})^{-\beta}$.

See R script

Expectation of power-laws

$$E[X] = \int_{x_{min}}^{\infty} xp(x)dx = C \int_{x_{min}}^{\infty} x^{-\alpha+1}dx = \frac{C}{2-\alpha}\left[x^{-\alpha+2}\right]_{x_{min}}^{\infty}$$

- Finite only for $\alpha > 2$:

$$E[X] = \frac{\alpha - 1}{\alpha - 2} x_{min}$$

  - For $1 < \alpha \leq 2$, there is no expectation: the mean of a sample data has no reliable value!
- $Var(X)$ finite only for $\alpha > 3$!
  - For $2 < \alpha \leq 3$, the sample variance of a dataset has no reliable value!

Discrete power-law

Discrete Power-law

A discrete random variable X has the *power-law distribution*, if for some $\alpha > 1$ its p.m.f. function is given by

$$p(k) = C \cdot k^{-\alpha} \quad \text{for } k = k_{min}, k_{min} + 1, \ldots$$

We denote this distribution by $Pow(k_{min}, \alpha)$.

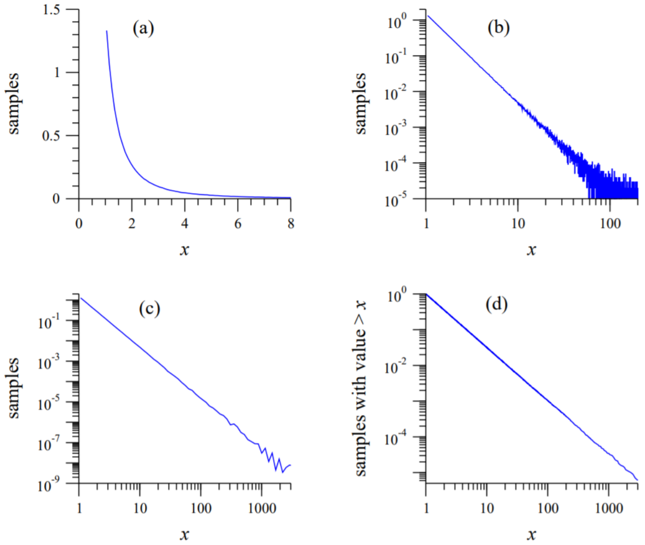
- Population of cities, number of books sold, number of citations, etc.
- Since $1 = \sum_{k=k_{min}}^{\infty} Ck^{-\alpha}$, we have

$$C = \frac{1}{\sum_{k=k_{min}}^{\infty} k^{-\alpha}} = \frac{1}{\zeta(\alpha, k_{min})}$$

  where $\zeta(\alpha, k_{min}) = \sum_{k=k_{min}}^{\infty} k^{-\alpha}$ *[Hurwitz zeta-function]*
- $\zeta(\alpha) = \zeta(\alpha, 1) \sum_{k=1}^{\infty} k^{-\alpha}$ *[Riemann zeta-function]*

See R script

Logarithmic binning vs CCDF



Zipf's law

Zipf's law

A discrete random variable X has the *Zipf's law distribution*, if for some $\alpha > 1$ its p.m.f. function is given by

$$p(r) = C \cdot r^{-\alpha} \quad \text{for } r = 1, 2, \ldots, N$$

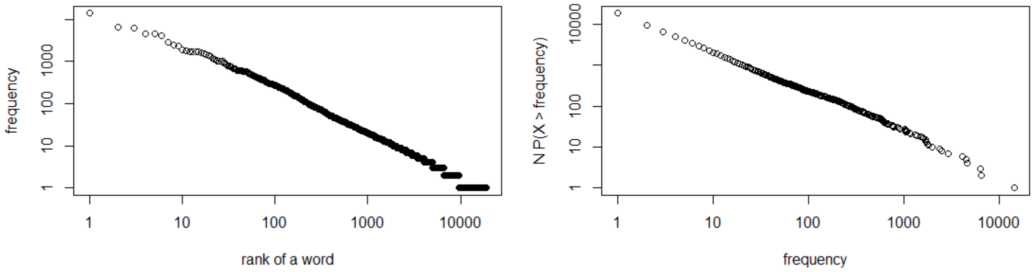
We denote this distribution by $\mathit{Zipf}(\alpha)$.

- Since $\sum_{r=1}^{N} Cr^{-\alpha} = 1$:

$$C = \frac{1}{\sum_{r=1}^{N} r^{-\alpha}} = \frac{1}{\zeta(\alpha) - \zeta(\alpha, N+1)}$$

- Read $p(r)$ as the probability of an event based on the "rank of" the event
    - e.g., prob. of occurrence of a given word in a book given the word rank, prob. of occurrence of a person of a given city given the city rank
    - If V the total number of words/inhabitants, $V \cdot p(r)$ is the frequency/population of the word/city of rank r. Alternatively, if v is the population of the city $p(r) = {}^{v}/V$

Zipf's law



Left Frequency of words based on rank *[Zipf's law]*

Right Number of words with a given minimum frequency *[CCDF of a Power-law]*

From power-law to Zipf's law

- $X \sim Pow(x_{min}, \alpha)$, e.g., population of a city
- $p(k) = Ck^{-\alpha}$, e.g., probability of having k inhabitants
- $P(X > k) = k^{-\alpha+1}$, e.g., probability of k or more inhabitants
- Let N be the number of cities and V the total population in all cities
- $r = N \cdot P(X > k) = N \cdot k^{-\alpha+1}$ is how many cities have k or more inhabitants
  - i.e., the rank of a city given its population
- Hence $r \propto k^{-\alpha+1}$ implies: *[$\propto$ reads "proportional to" up to multiplicative/additive constants]*

$$p(r) = \frac{k}{V} \propto k \propto r^{-\beta} \quad \text{for } \beta = \frac{1}{\alpha - 1}$$

- The r^{th} most populated city has population proportional to $r^{-\beta}$

See R script