CSE 591: Data Science
Steven Skiena
Stony Brook University

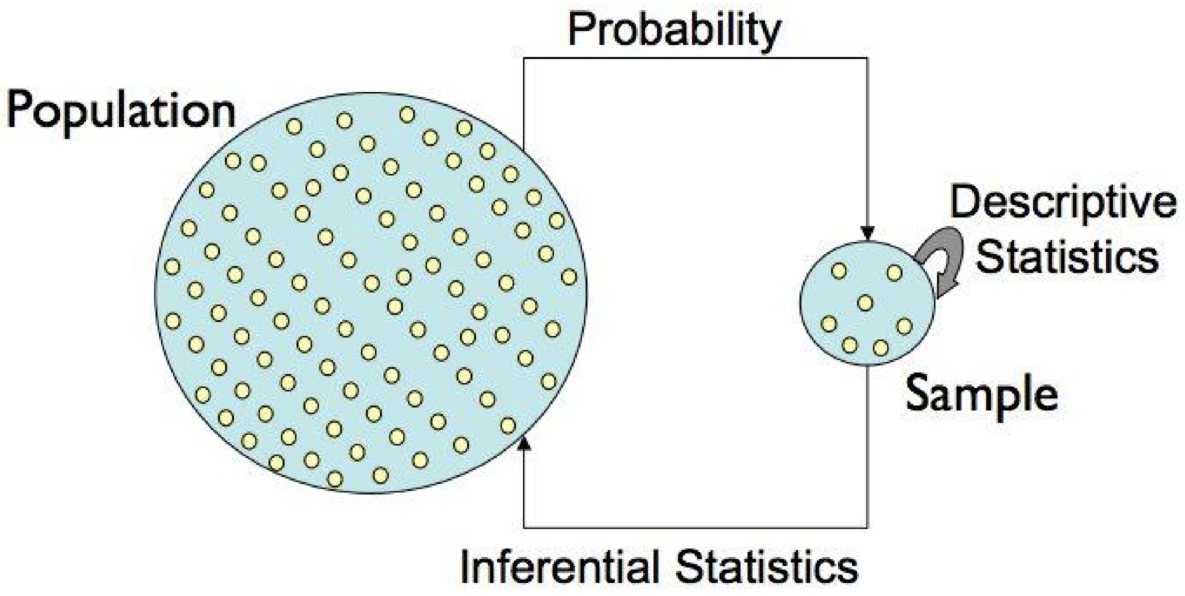
Lecture 9: Statistical Distributions

Statistics and Data Science

“A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.”

- Josh Blumenstock (Univ. of Washington)

The Central Dogma of Statistics



Statistical Data Distributions

Every observed random variable has a particular frequency/probability distribution.

Some distributions occur often in practice/theory:

- The Binomial Distribution
- The Normal Distribution
- The Poisson Distribution
- The Power Law Distribution

Significance of Classical Distributions

Classical probability distributions arise often in practice, so look out for them.

Closed-form formulas and special statistical tests often exist for particular distributions.

However, your observed data does not necessarily come from a particular distribution just because the shape looks similar.

Binomial Distributions

Experiments consist of *n identical, independent* trials which have two possible outcomes, with probabilities *p* and *(1-p)* like heads or tails.

$$P\{X = x\} = \binom{n}{x} p^{x} (1 - p)^{n-x}$$

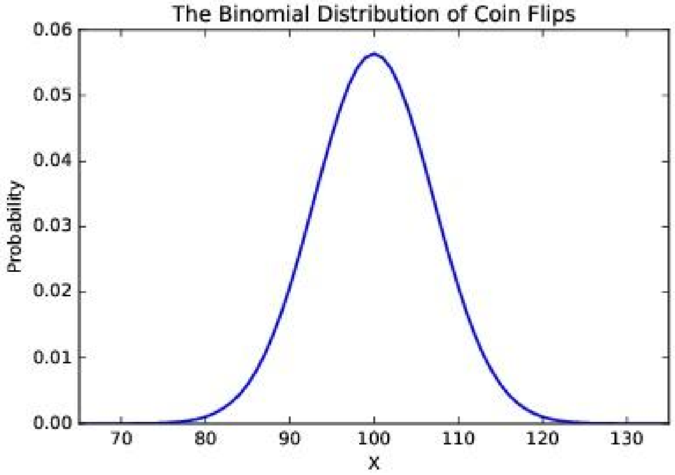
The observed season batting averages of a *p=0.300* hitter were drawn from a binomial distribution.

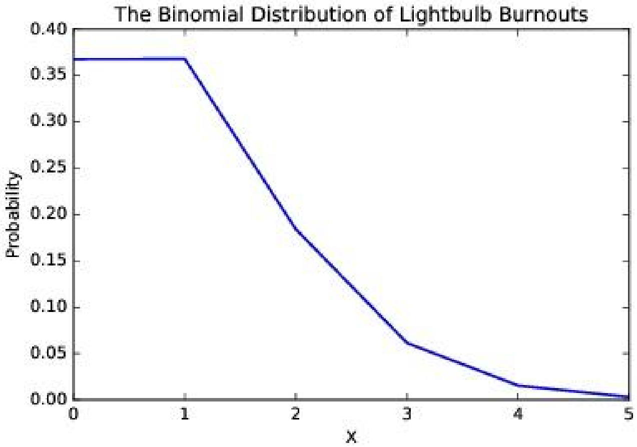
Properties of Binomial Distributions

Discrete, but bell (or half-bell) shaped

Coin flips: p=0.5 n=100

Lightbulb burnouts: p=0.001 n=1000





The distribution is a function of n and p.

The Normal Distribution

The bell-shaped distribution of height, IQ, etc.

Completely parameterized by mean and standard deviation:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

Not all bell-shaped distributions are normal but it is generally a reasonable start.

Properties of the Normal Distribution

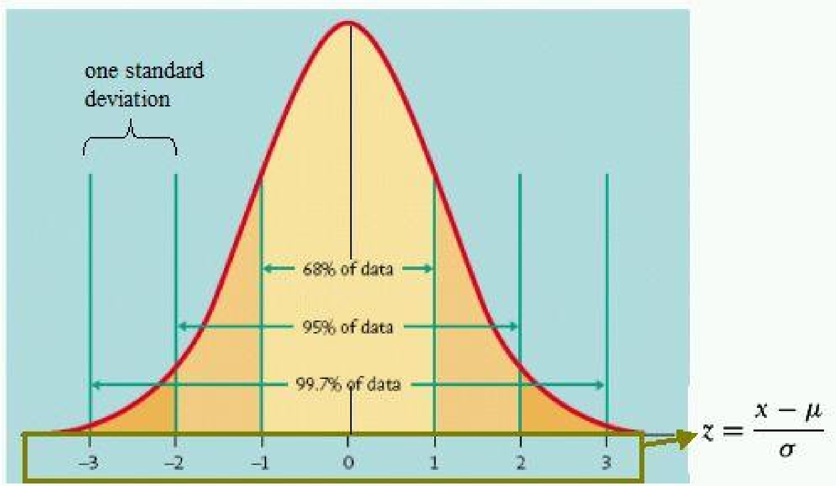
- It is a generalization of the binomial distribution where $n \to \infty$
- Instead of *n* and *p*, the parameters are the mean *mu* and standard deviation *sigma*.
- It really **is** bell-shaped since *x* is continuous and goes infinitely in each direction.
- The sum of normally distributed variables is normal.

Interpreting the Normal Distribution

Tight bounds on probability follow for Z-scores from normally distributed random variables:

IQ is normally distributed, with mean 100 and standard deviation 15.

Thus about 2.5% of people have IQs above 130.

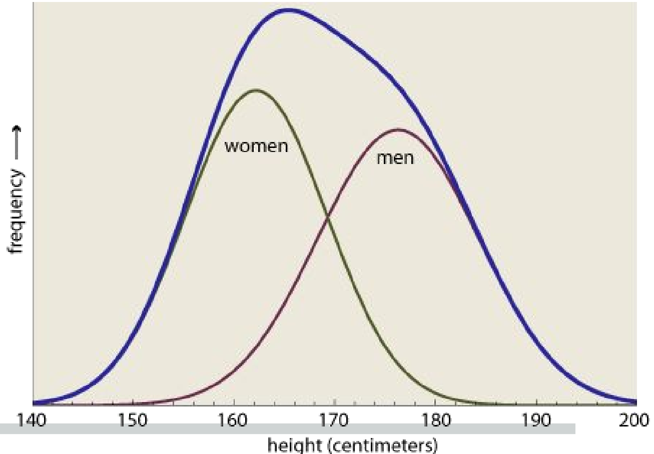


What’s not Normal?

Not all bell-shaped distributions are normal (i.e. stock returns are log normal with fat tails).

Mixtures of normal distributions are not normal, like full population heights.

Statistical tests exist to establish whether data is drawn from a normal distribution, but populations are generally mixtures of multiple distributions: height, weight, IQ



Lifespan Distributions

If your chance of surviving any given day is probability *p*, what is your lifespan distribution?

A lifespan of *n* days means dying for the first time on day *n*, so

$$Pr(n) = p^{n-1}(1-p)$$

Lightbulb life spans are better modeled with such a distribution, not dead bulbs per 1000 hours.

The Poisson Distribution

The Poisson distribution measures the frequency of intervals between rare events.

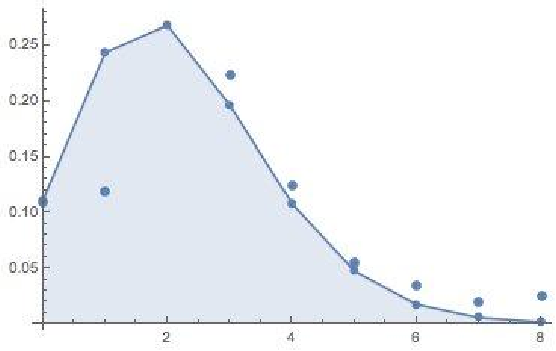
$$Pr(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

Instead of event probability *p*, the distribution is parameterized by mean *mu*, but this is equivallent because $\mu = \sum_{k=0}^{\infty} k \cdot Pr(k)$

Distribution of Kids per Family

The average U.S. family has 2.2 kids, but how are they distributed?

If families repeatedly decide whether to have any more children with fixed probability *p* we get a Poisson distribution:



Power Law Distributions

Power laws are defined $p(x) = c x^{-a}$, for exponent a and normalization constant c.

They do not cluster around a mean like a normal distribution, instead having very large values rarely but consistently.

They define 80-20 rules: 20% of the X get 80% of the Y.

City Population Yield Power Laws

The average big US city has population 165,719.   Even with a huge standard deviation of 410,730, the biggest city under a normal distribution should be Indianapolis (780K).

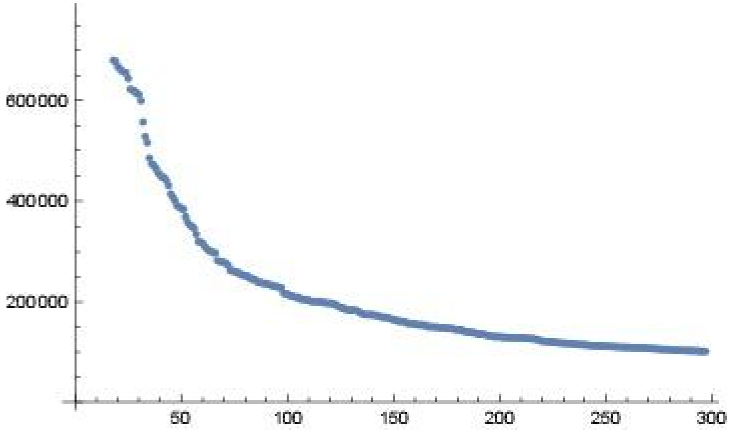
New York city had 8,008,278 people in the 2000 census.

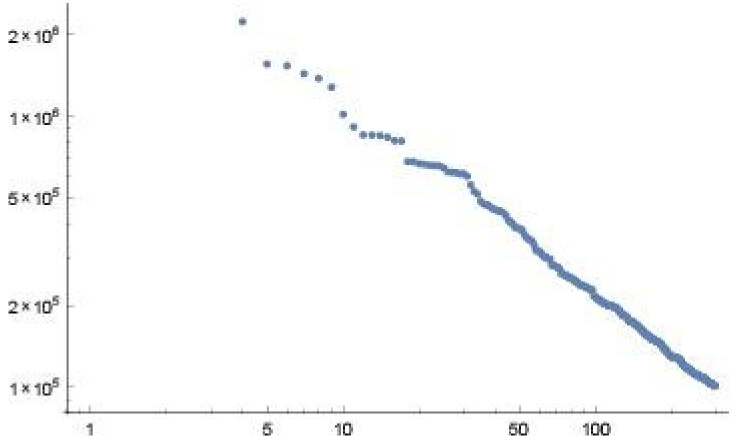
Power laws arise when the rich get richer.

Linear and Log-Log Plots for City Pop

Straight lines on log-log plots say power law.

The biggest values are out of scale on linear plots.





Wealth Yields Power Laws

1 Bill Gates has $80 billion.

5 Hyperbillionaries have $40 billion each.

25 SuperBillionaries have $20 billion each.

125 MultiBillionaries have $10 billion each.

625 Billionaries have $5 billion each.

Power law: as you multiply the value by *x*, you divide the number of people by *y*.

Definitions of Power Laws

For a power law distributed variable X,

$$P(X = x) = cx^{-a}$$

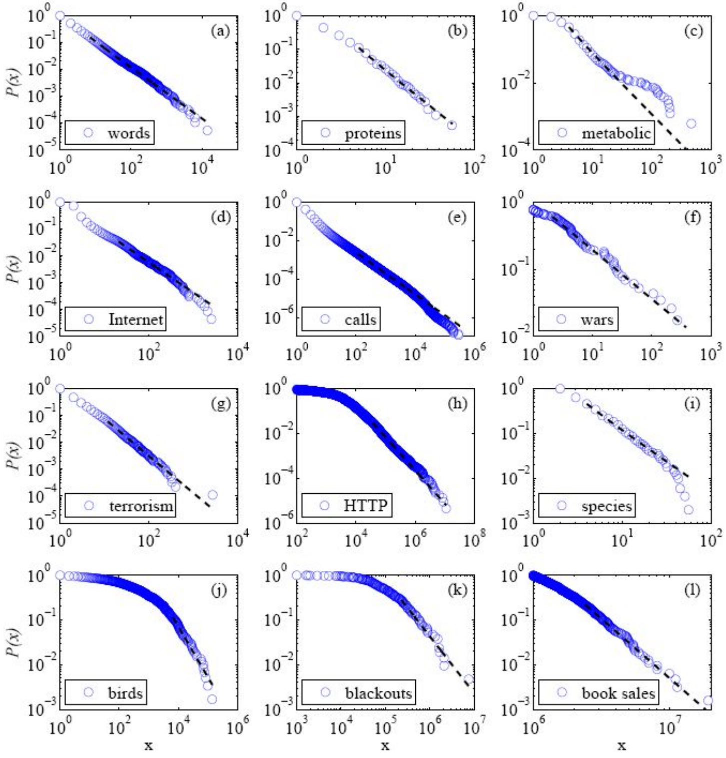
The constant *c* is unimportant: for a given *a* this constant *c* ensures the probability sums to 1.

Doubling x (to 2x) reduces the probability by a factor of 2^a, so larger values keep getting rarer at steady, non-decreasing rate.

Many Distributions are Power Laws

- Internet sites with x inlinks.
- Frequency of earthquakes at x on the Richter scale
- Words used with a relative frequency of x
- Wars which kill x people

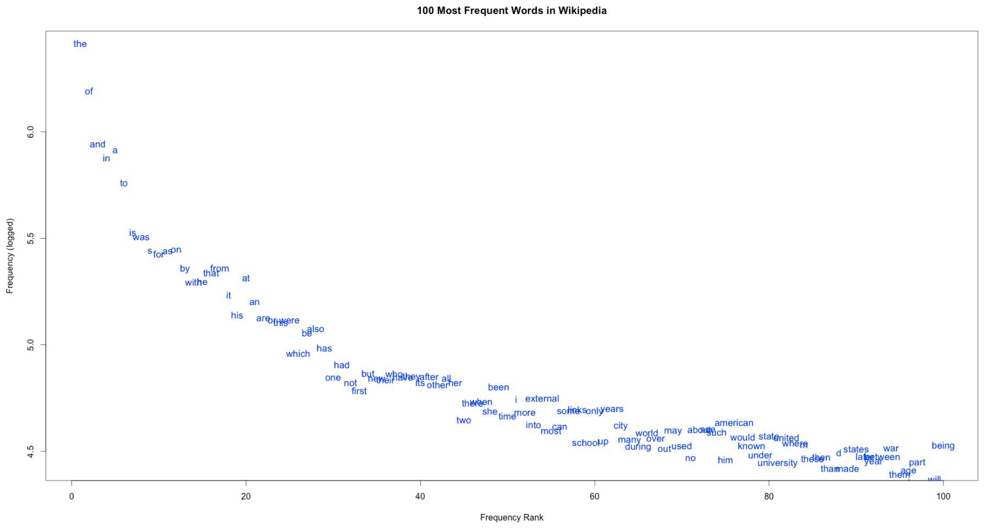
Power laws show as straight lines on log value, log frequency plots.



Word Frequencies and Zipf’s Law

Zipf’s law states that the kth most popular word is used 1/kth as often as the most popular word.

Zipf’s law is a power law for *a=1*, so a word of rank *2x* have half the frequency of rank *x.*



Properties of Power Laws

- The mean does not make sense. Bill Gates adds about $250 to the US mean wealth.
- The standard deviation does not make sense, typically much larger than the mean.
- The median better captures the bulk of the distribution.
- The distribution is *scale invariant*, meaning zoomed in regions look like the whole plot.