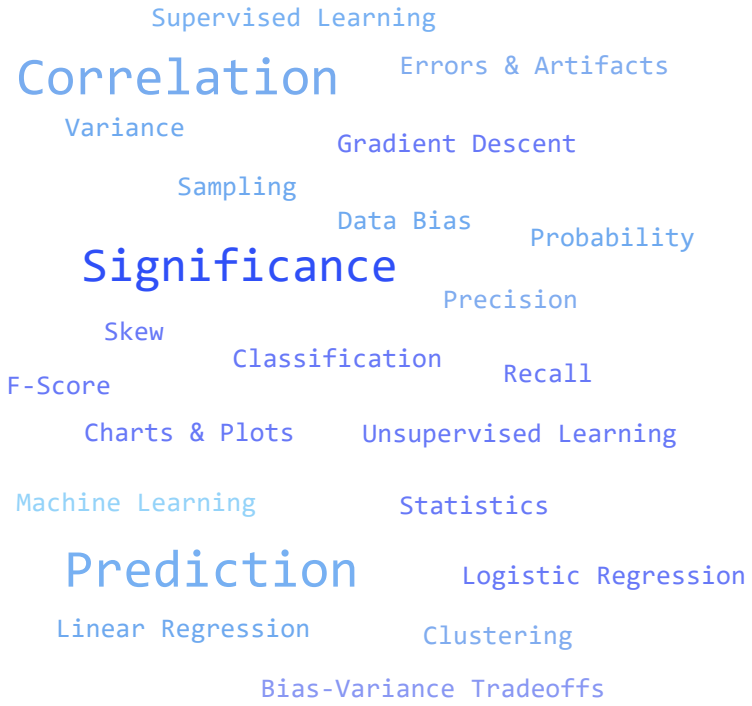


Data Science 1:
Introduction to Data Science

Statistical Distributions & Significance

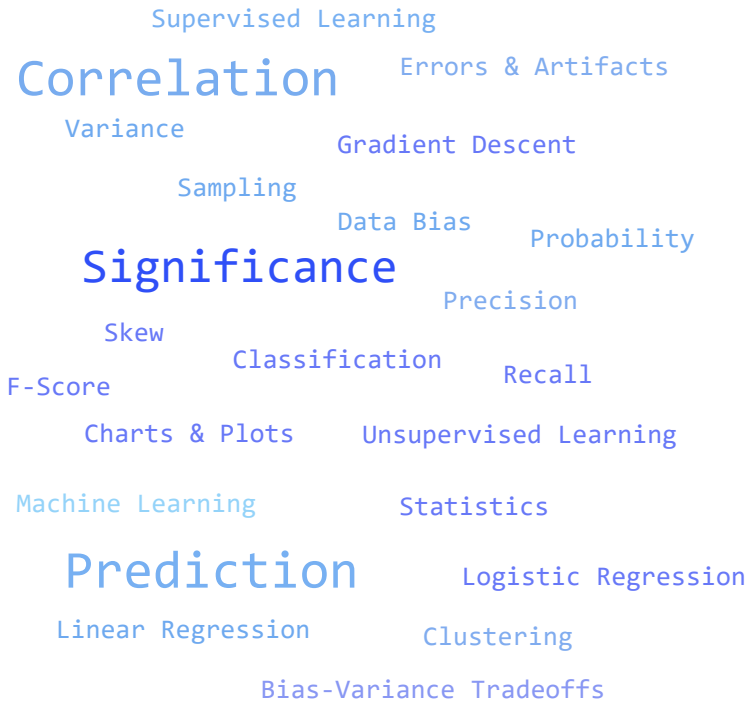
Winter 2025

Wolfram Wingerath, Jannik Schröder

Department for Computing Science
Data Science / Information Systems

Lecture slides based on content from “The Data Science Design Manual” (Steven Skiena, 2017) and associated course materials generously made available online by the author at https://www3.cs.stonybrook.edu/~skiena/data-manual/.

Special thanks to Professor Skiena for sharing these valuable teaching resources!



Data Science 1:
Introduction to Data Science

Statistical Distributions & Significance

Winter 2025

Wolfram Wingerath, Jannik Schröder

Department for Computing Science
Data Science / Information Systems

Semester Schedule

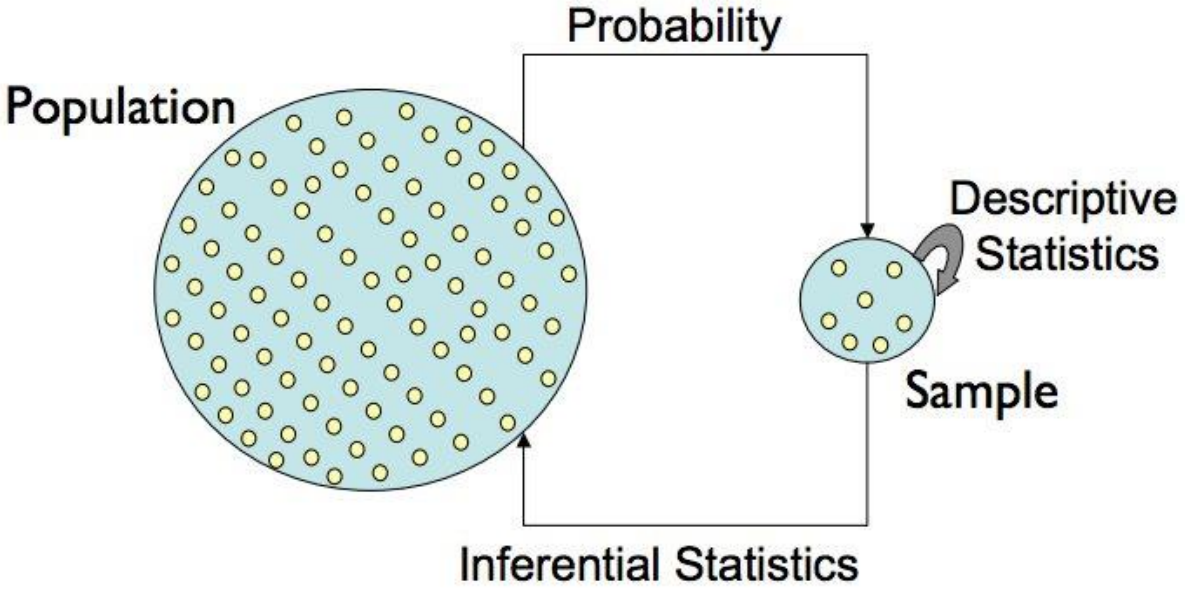
CW 42	14. Oct	Lecture	1	Orga & Intro	1-26
CW 43	21. / 23. Oct	Lecture + Exercises	2	Probability, Statistics & Correlation	27-56
CW 44	28. Oct	Lecture	3	Data Munging, Cleaning & Bias	57-94 / “Invisible Women”
CW 45	04. / 06. Nov	Lecture + Exercises	4	Scores & Rankings	95-120
CW 46	11. Nov	Lecture	5	Statistical Distributions & Significance	121-154
CW 47	18. / 20. Nov	Lecture + Exercises	6	Building & Evaluating Models	201-236
CW 48	25. Nov	Guest Lecture	7	Data Visualization	155-200
CW 49	02. / 04. Dec	Lecture + Exercises	8	Intro to Machine Learning	351-390
CW 50	09. Dec	Lecture	9	Linear Algebra	237-266
CW 51	16. / 18. Dec	Lecture + Exercises	10	Linear Regression & Gradient Descent	267-288
CW 02	06. Jan	Lecture	11	Logistic Regression & Classification	289-302
CW 03	13. / 15. Jan	Lecture + Exercises	12	Nearest Neighbor Methods & Clustering	303-350
CW 04	20. Jan	Lecture	13	Data Science in the Wild	391-426
CW 05	27. / 29. Jan	Lecture + Exercises	14	Q&A / Feedback	
CW 06	03. / 04. Feb	Oral Exams (Block 1)		Preparation in our last session („Oral Exam Briefing“)	
CW 13	24. / 25. Mar	Oral Exams (Block 2)			

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



Joshua Akey et.al.: Genome 560, Lecture Slides (2008)

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.

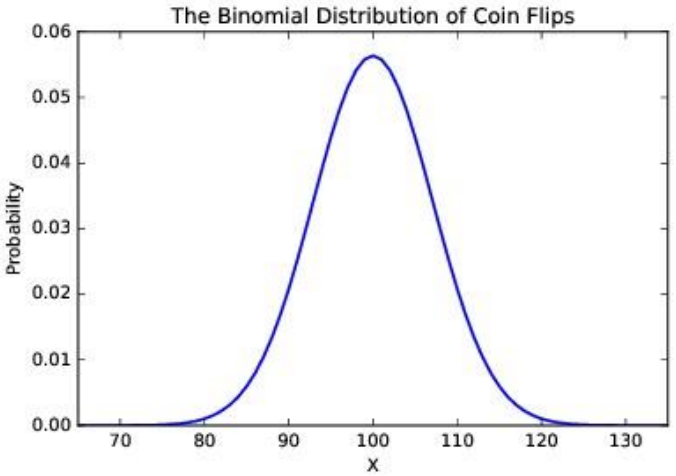
$$P\{X = x\} = \binom{n}{x} p^x (1-p)^{n-x}$$

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

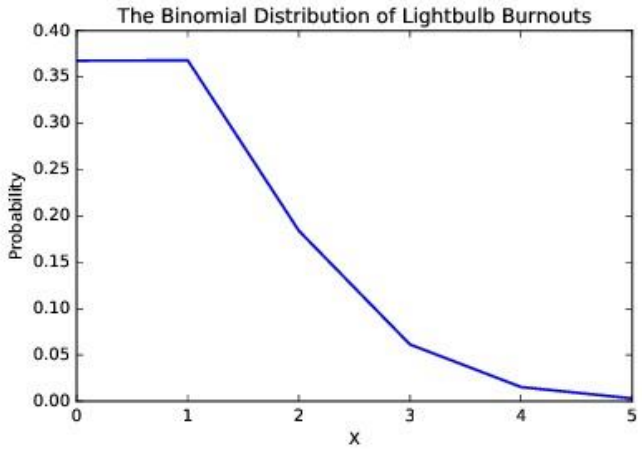
Properties of Binomial Distributions

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.

Variance of the Binomial Distribution

The closed form is: $\sigma = \sqrt{np(1-p)}$

The mode, mean, and median is *np* for integers.

The probability of getting exactly $\frac{n}{2}$ heads is $O(\frac{1}{\sqrt{n}})$, so it is actually quite low.

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

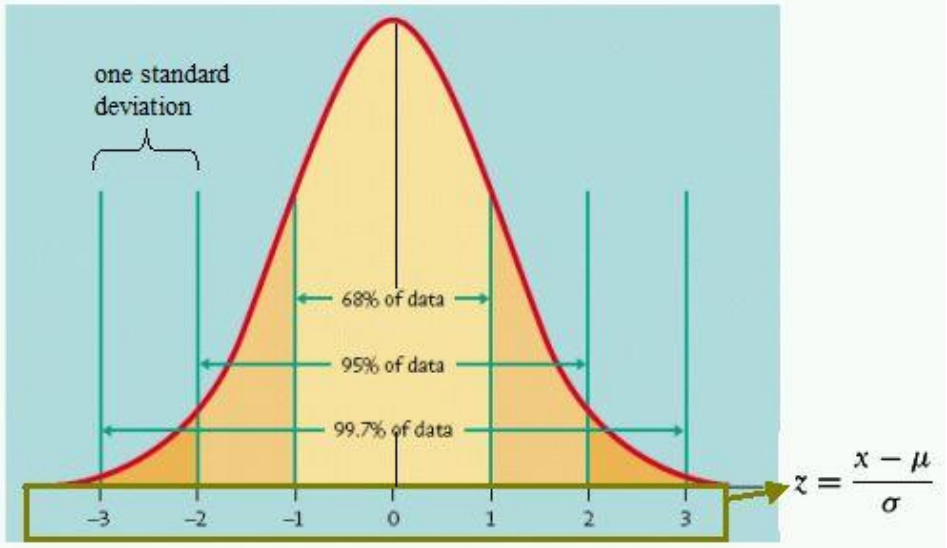
- It is a generalization of the binomial distribution where $n \to \infty$
- Instead of *n* and *p*, the parameters are the mean μ (*mu)* and standard deviation σ (*sigma)*.
- It really **is** bell-shaped since *x* is continuous and goes infinitely in each direction.
- The sum of independent normally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

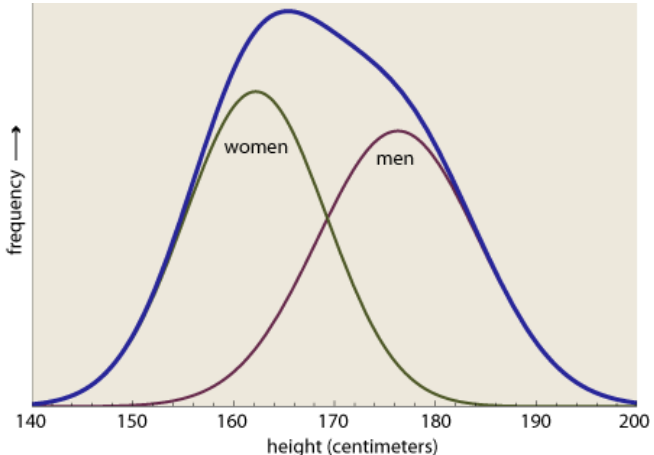


What's not Normal?

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, ...



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.

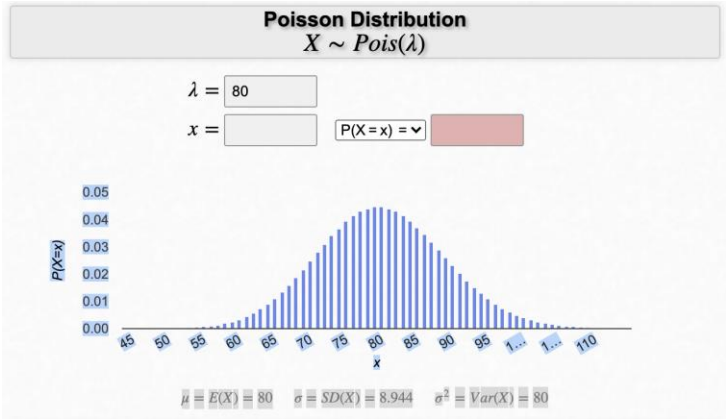
$$Pr(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

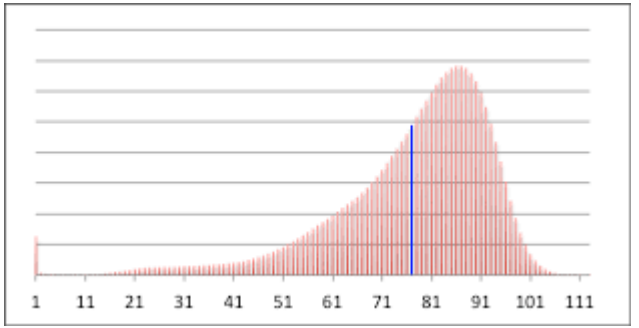
Instead of event probability *p*, the distribution is parameterized by mean *mu*, but this is equivalent because $\mu = \sum_{k=0}^{\infty} k \cdot Pr(k)$

Distribution of Human Lifespans

From an average lifespan = 80, the Poisson distribution looks kind of reasonable.

But is the probability of dying each day constant over all ages?

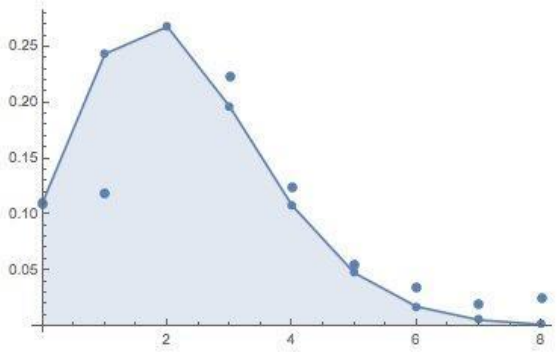




Distribution of Kids per Family

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 $p(x) = cx^{-a}$, 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).

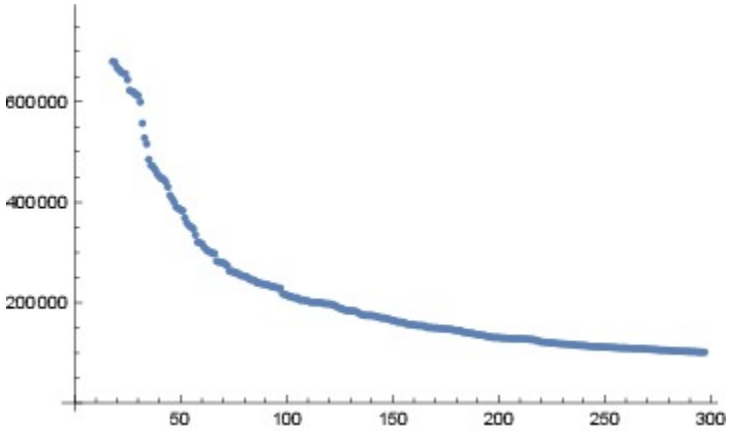
New York city had 8,008,278 people in the 2000 census.

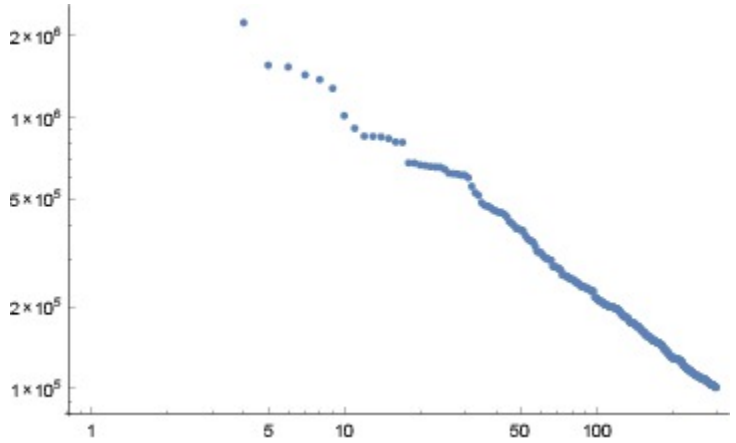
Power laws arise when the rich get richer.

Linear & 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,

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

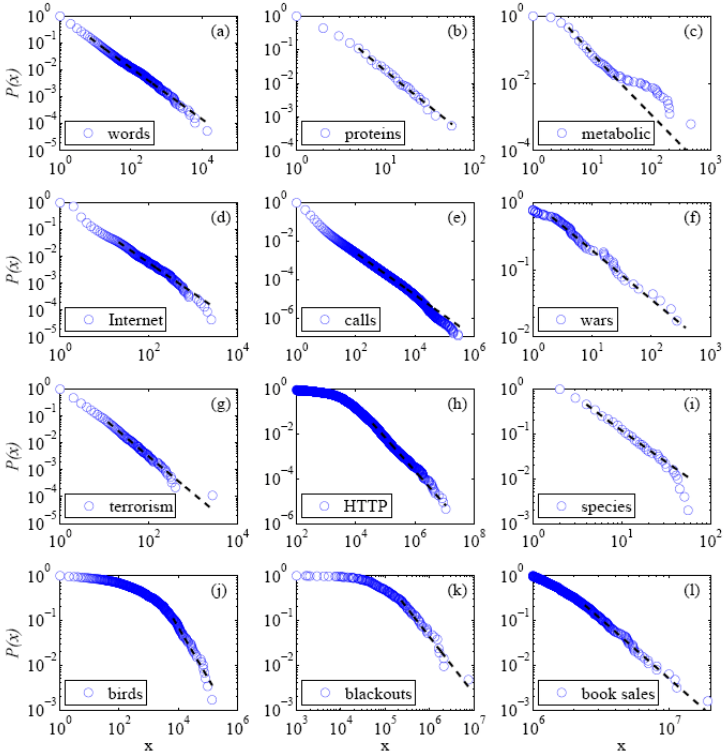
The constant *c* is unimportant: for a given *a* this constant *c* ensures the probability sums to 1.

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

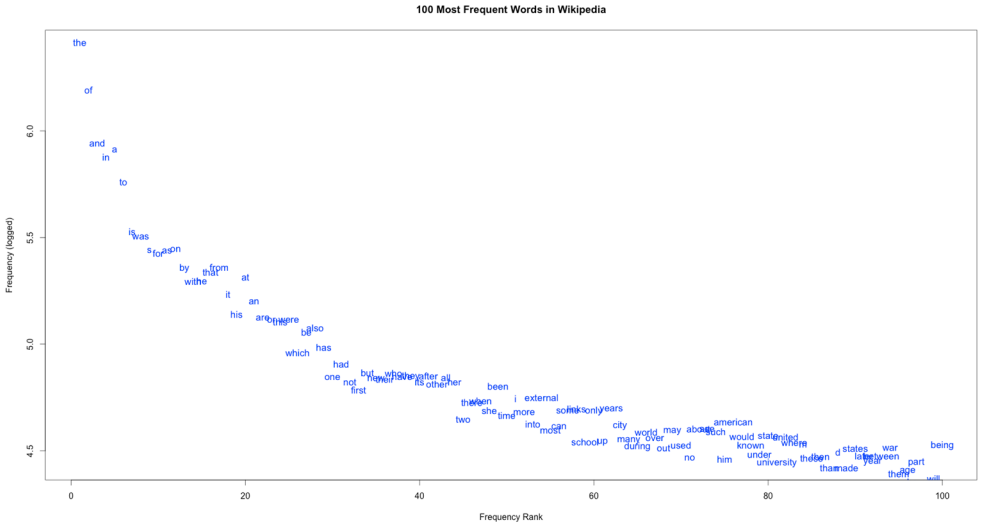
Power laws show as straight lines on log value, log frequency plots.



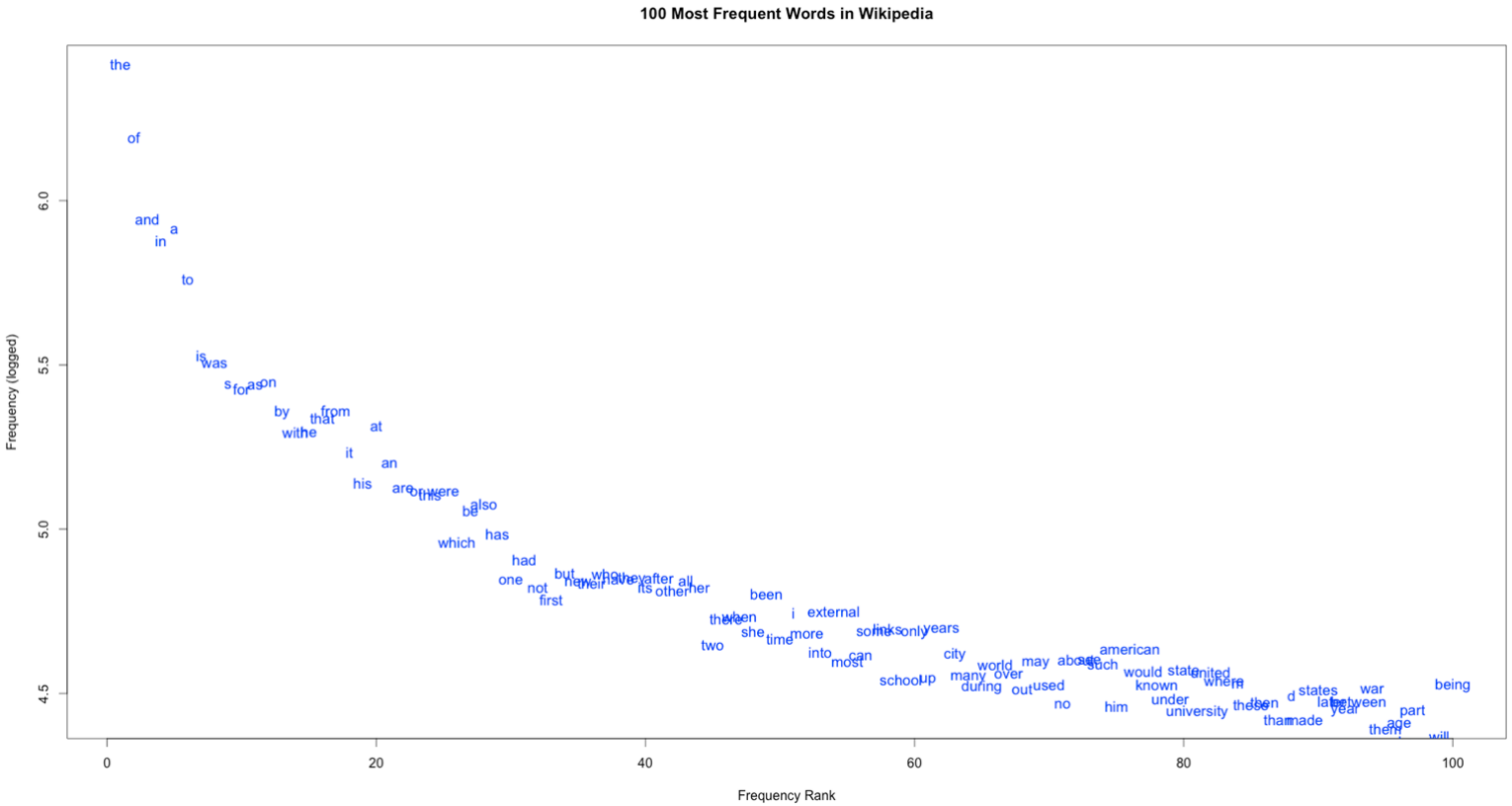
Word Frequencies and Zipf's Law

Zipf's law states that the kth most popular word is used $1/k^{th}$ as often as the most popular word.

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

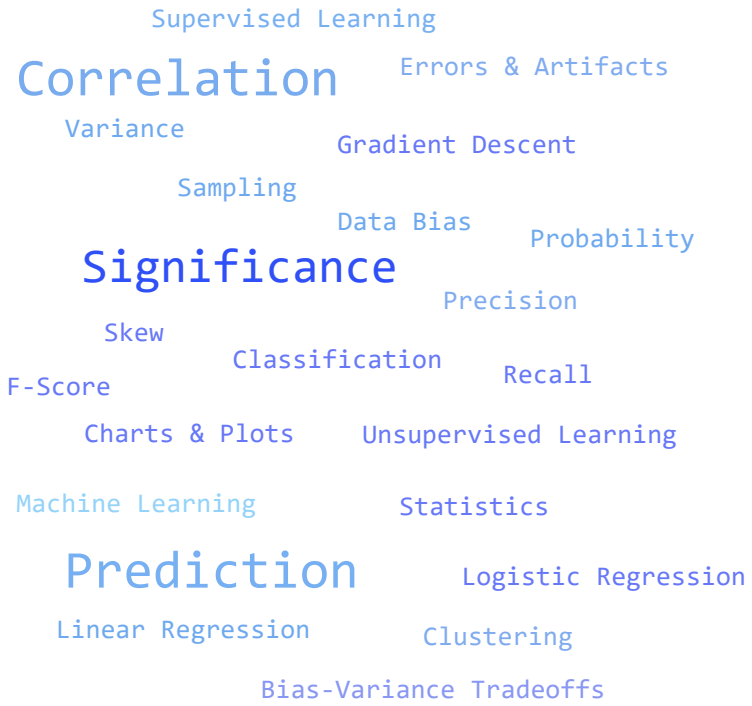


Word Frequencies and Zipf's Law



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.



Data Science 1:
Introduction to Data Science

Statistical Distributions & Significance

Winter 2025

Wolfram Wingerath, Jannik Schröder

Department for Computing Science
Data Science / Information Systems

Talking to Statisticians

Statisticians are primarily concerned with whether observations on data are significant.

Data miners are primarily concerned with whether their observations are interesting.

When is an Observation Meaningful?

Computational analysis readily finds patterns and correlations in large data sets.

But when is a pattern significant?

Sufficiently strong correlations on large data sets may seem *obviously* significant, but often the effects are more subtle.

Medical Statistics

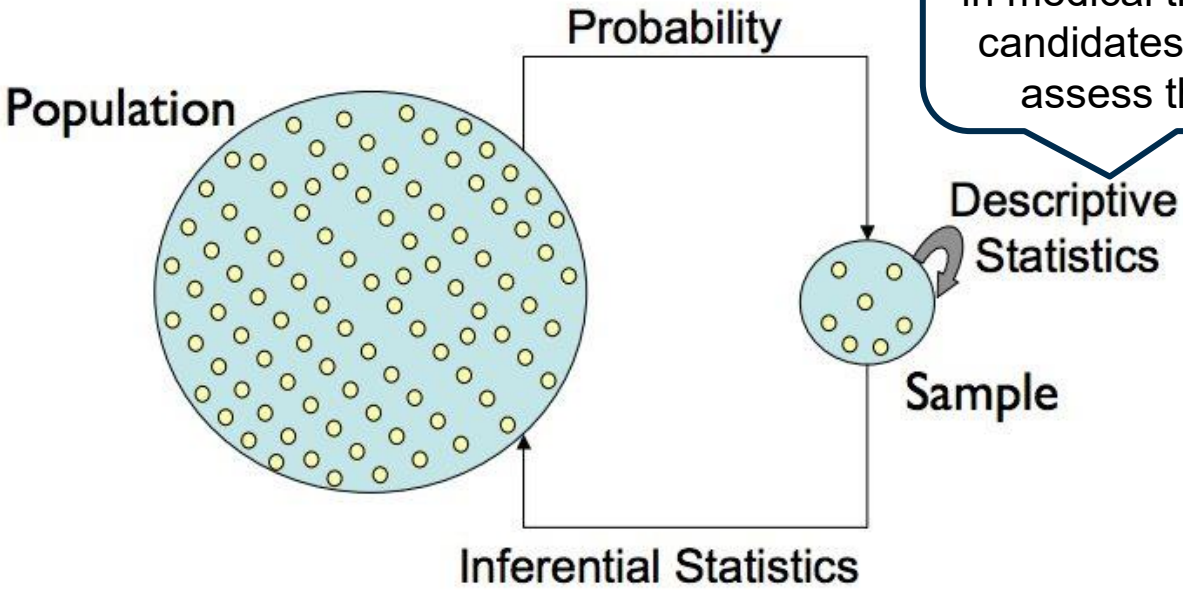
Evaluating the efficacy of drug treatments is a classically difficult problem.

Drug A cured 19 of 34 patients. Drug B cured 14 of 21 patients. Is B better than A?

FDA approval of new drugs rests on such trials/analysis, and can add/subtract billions from the value of drug companies.

Remember: Sample vs. Population

People may die for collecting your sample (e.g. in medical trials). How many candidates do you need to assess the new drug?



Significance and Classification

In building a classifier to distinguish between two classes, it pays to know whether input variables show a real difference among classes.

Is the length distribution of spam different than that of real mail?

Comparing Population Means

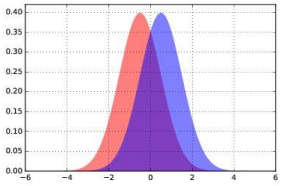
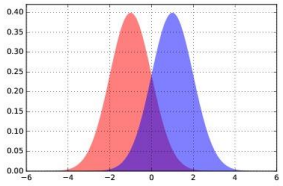
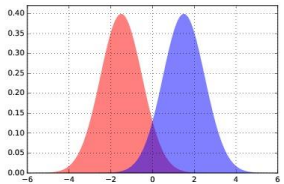
The T-test evaluates whether the population means of two samples are different.

Sample the IQs of 20 gamers and 20 non-gamers. Is one group smarter on average?

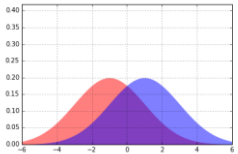
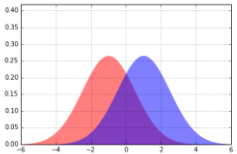
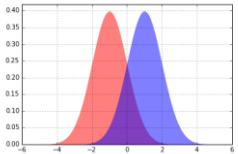
Certainly the sample means will differ, but is this difference significant?

Differences in Distributions

It becomes easier to distinguish two distributions as the means move apart...

... or the variance decreases:

The T-Test

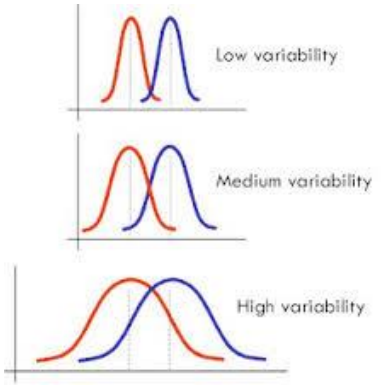
Two means differ significantly if:

- The mean difference is relatively large
- The standard deviations are small enough
- The samples are large enough

Welch's t-statistic is: $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$

where s^2 is the sample variance.

Significance is looked up in a table.



Why Significance Tests Can Work

Statistical tests seem particularly opaque (e.g. look up numbers from table), but come from ideas like:

- Probabilities of samples drawn from distributions with given mean and std. dev.
- Bayes theorem converts Pr(data|distribution) to Pr(distribution|data)

The Kolmogorov-Smirnov Test

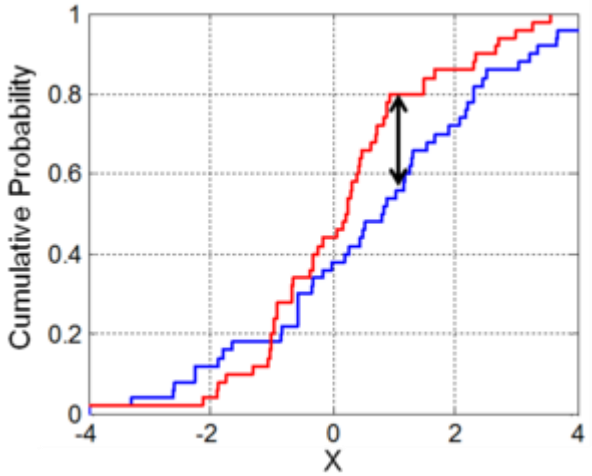
This test measures whether two samples are drawn from same distribution by the maximum difference in their cdf.

The distributions differ if:

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|,$$

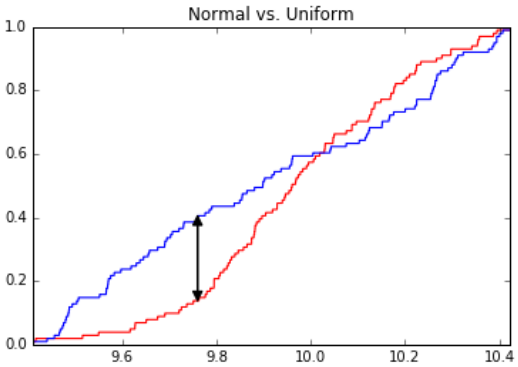
and $D_{n,n'} > c(\alpha)\sqrt{\frac{n+n'}{nn'}}.$

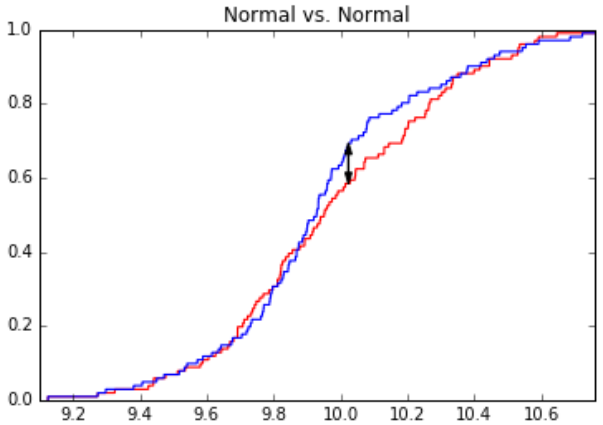
at a significance of alpha.



Normality Testing

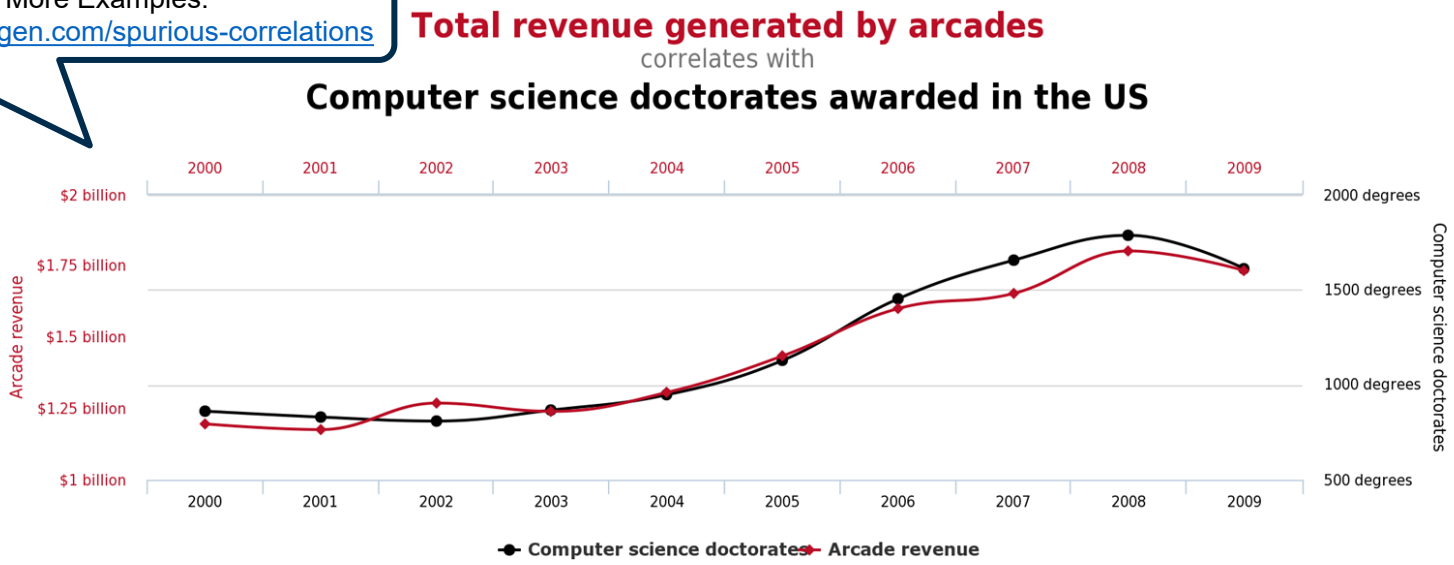
We can perform the KS-test where one distribution is sampled from the theoretical distribution:





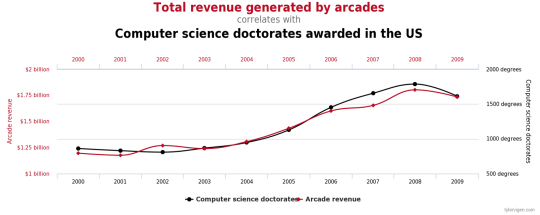
The Bonferroni Correction

More Examples:
https://tylervigen.com/spurious-correlations



If you compare enough distributions with another, you will find is strong correlations somewhere.

The Bonferroni Correction



A statistical significance of 0.05 means there is a probability 1/20 this result came by chance.

Thus “fishing expeditions” which test millions of hypotheses must be held to higher standards!

In testing *n* hypotheses, one must rise to a level of α/n to be considered significant at the level of *alpha*.

The Significance of Significance

For large enough sample sizes, extremely small differences can register as highly significant.

Significance measures the confidence there is a difference between distributions, not the effect size or importance/magnitude of the difference.

Measures of Effect Size

- *Pearson correlation coefficient:* small effects start at 0.2, medium effects at 0.5, large effects at 0.8
- *Percentage of overlap between distributions:* small effects start at 53%, medium effects at 67%, large effects at 85%
- *Cohen's d* $d = (|\mu - \mu')/\sigma)$: small >0.2, medium > 0.5, large > 0.8

Bootstrapping P-values

Traditional statistical tests evaluate whether two samples came from the same distribution.

Many have subtleties (e.g. one- vs. two-sided tests, distributional assumptions, etc.)

Permutation tests allow a more general, more computationally idiot-proof way to establish significance.

Permutation Tests

If your hypothesis is true, then randomly shuffled data sets should not look like real data.

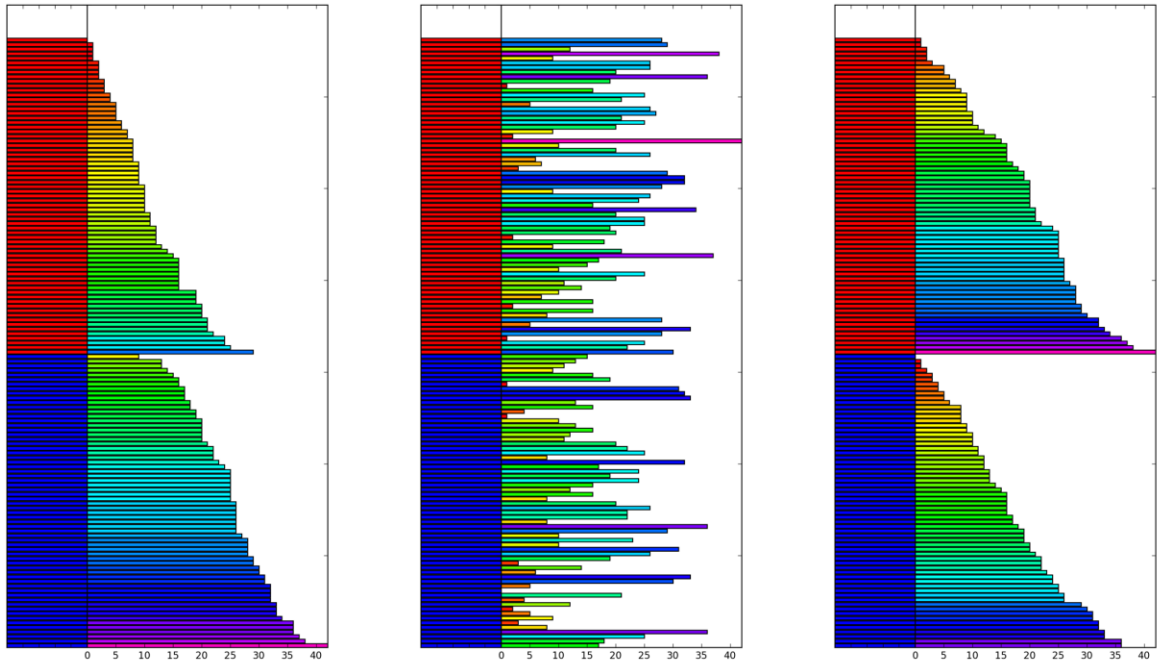
The ranking of the real test statistic among the shuffled test statistics gives a p-value.

You need statistic on your model you believe is interesting, e.g. correlation, std. error, or size.

Permutation Test (Gender Relevant?)

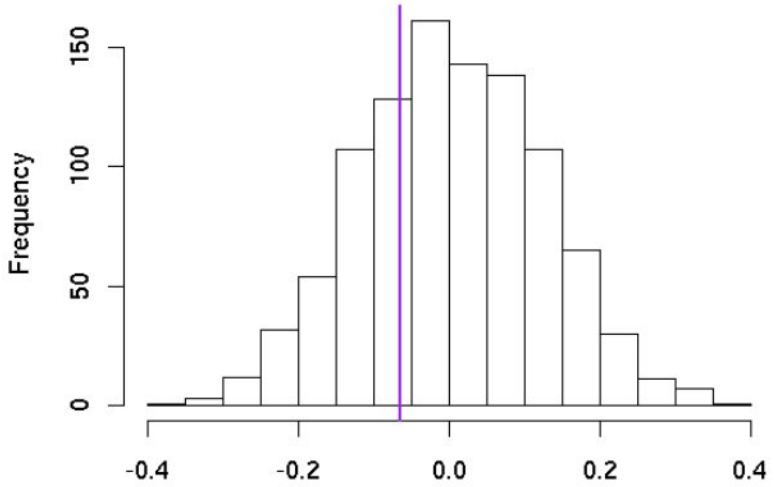
Heights here coded by bar length and color

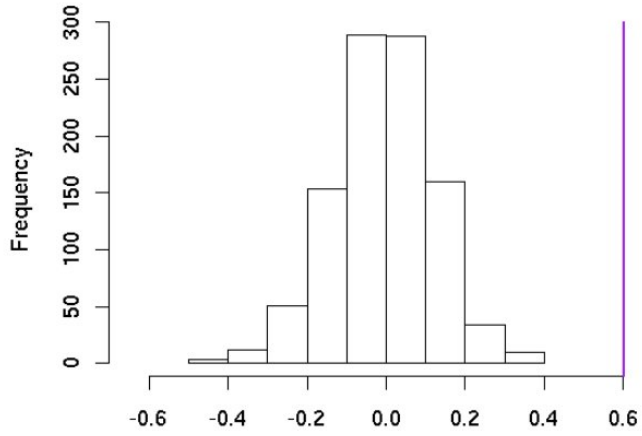
The random permutation (c/r) shows less height difference by gender than the original data (l).



Significance of a Permutation Test

The rank of the real data among the random permutations determines significance:





Performing Permutation Tests

The more permutations you try (at least 1000), the more impressive your significance can be.

Typically we permute the values of fields across records or time-points within a record. Keep comparisons apples-to-apples.

If your model shows decent performance trained on random data, you have a problem.

Permutation Test Caveat!

Permutation tests give you the probability of your data given your hypothesis.

This is not the same as the probability of your hypothesis given your data, which is the traditional goal of significance testing.

The real strength of your conclusion does not infinitely increase with more permutations!

Constructing Random Permutations

Constructing truly random permutations is surprisingly subtle. Which algorithm is right?

```
for i = 1 to n do a[i] = i;
for i = 1 to n − 1 do swap[a[i], a[Random[i, n]];
```

or:

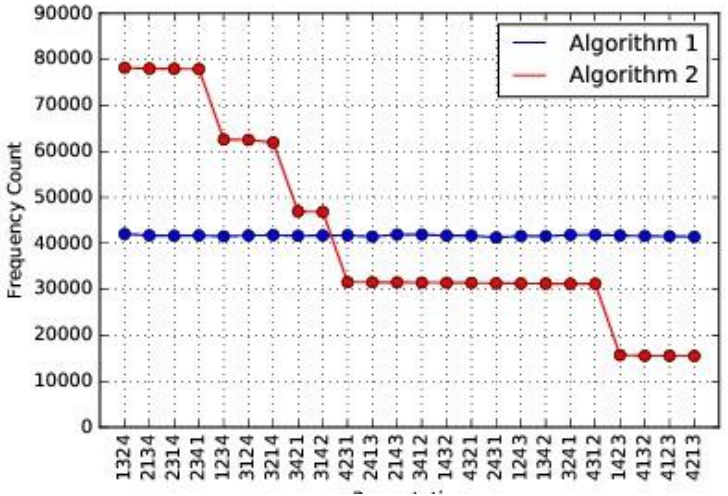
```
for i = 1 to n do a[i] = i;
for i = 1 to n − 1 do swap[a[i], a[Random[1, n]];
```

Yes, there is a difference

Experiments constructing 1 million random permutations shows that algorithm 1 is uniform, but algorithm 2 is not.

st. dev. 1 = 166.1

st. dev. 2 = 20,932.9



Why is it Uniform?

The first algorithm picks a random choice for the first position, then leaves it alone and recurs. It generates random permutations.

The second algorithm gives subsequent elements a better chance to end up first. The distribution is not uniform.

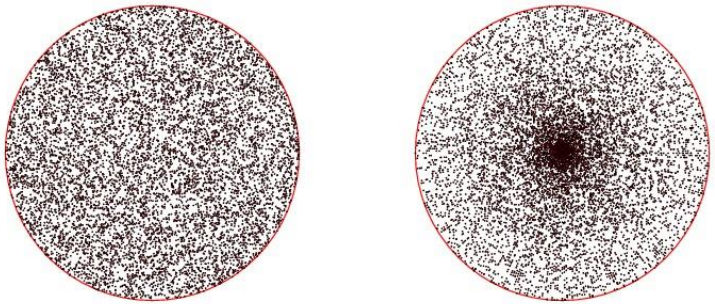
Moral: Random generation can be very subtle.

Sampling from Distributions

A common task is repeatedly drawing random samples from a given probability distribution.

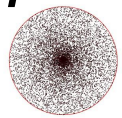
Give me an algorithm to draw uniformly random points from a circle:

The problem is more subtle than it looks.



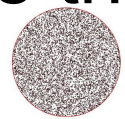
Drawing Points from a Circle

Each point in a circle is described by a radius *r* and angle *a*, but drawing them uniformly at random picks too many points near the center.



The inner half circle is smaller than the outer half!

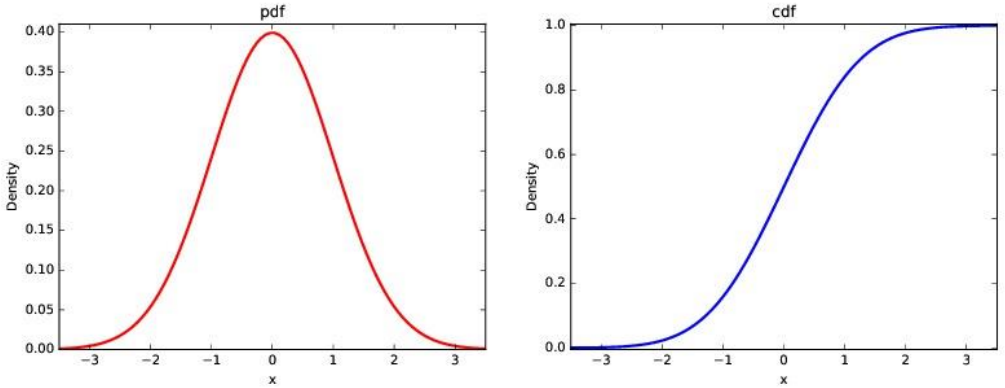
Independently sampling *x* and *y* give points uniform in the box, so discarding those outside the circle leaves a uniform distribution.



Inverse Transform Sampling

To sample from any probability distribution, convert it to its cumulative distribution (cdf).

Selecting a probability *p* in *[0,1]* now maps to a value in the cdf:



Dimaggio's Hitting Streak

One of baseball's most amazing records is Joe Dimaggio's 56-game hitting streak.

But how unusual is such a long streak in the context of his career?

He played 1736 games, with 2214 hits in 6821 at bats.

Thus he got a hit in roughly 79% of his games.

Monte Carlo Simulation

We can use random numbers to simulate when he got hits in over a synthetic “career”, and compute the length of the longest streak.

After simulating 100,000 Dimaggio career’s, we get a frequency distribution of longest streaks.

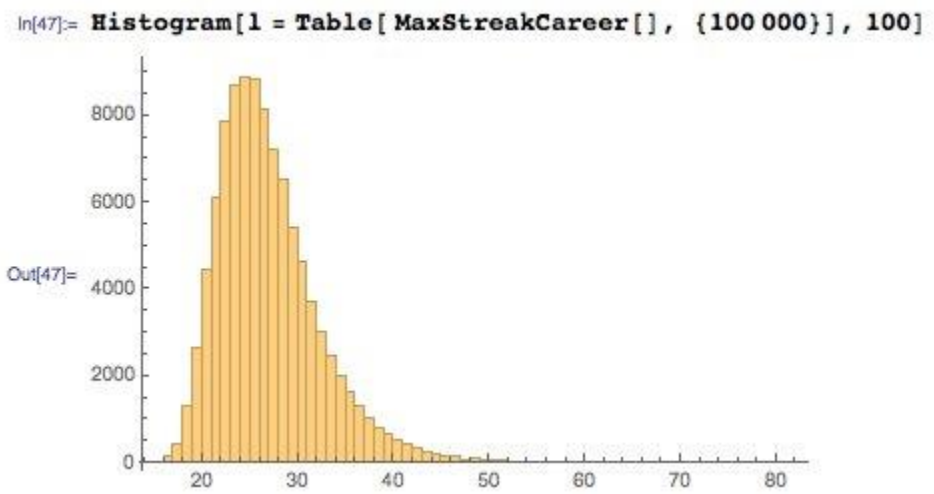
Simulation Results

In only 44/100000 simulated careers (1/2272) did he have a streak of at least 56 games.

Thus the length is quite out of line with what is expected from him, though he hit in 61 straight games in the minors.

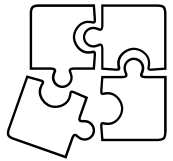
The second longest streak of any major league hitter is only 44 games, so it is out of line with everyone else as well.

Closed-form results could presumably follow from analyzing a Poisson distribution, but this requires more skill.



Wrapup:

Statistical Distr. & Significance



- Knowing the distribution of your data tells you a lot about your data!
- There are classical distributions (e.g. Binomial, Normal, Poisson, Power Law)
- Significance tests measure our confidence that an observation is not due to luck (e.g. T-Test)
- Sampling from a distribution is useful for analysis, but also easy to do wrong