DS 250: Data Analysis and Visualization

Lecture 6: Statistics for Data Scientists

Dr. Gagan Gupta

Slides based on “Data Science Design Manual” book by Dr. Steven Skiena

Announcements

- Assignment 1 due (8/21)
  - Please submit each playlist recommendation as part of your Notebook
  - Answer some specific questions
- Quiz 3 on 8/18. No extra time from now on…
- Project Feedback
  - Please reply to my feedback with your entire group
  - In some cases, I have suggested modifications
  - In other cases, start writing more details about your approach
  - 1 page proposal due (9/1)
- Office hours, Project discussion, difficulty with the course?
  - Please contact me on Piazza/email or talk to your class representatives

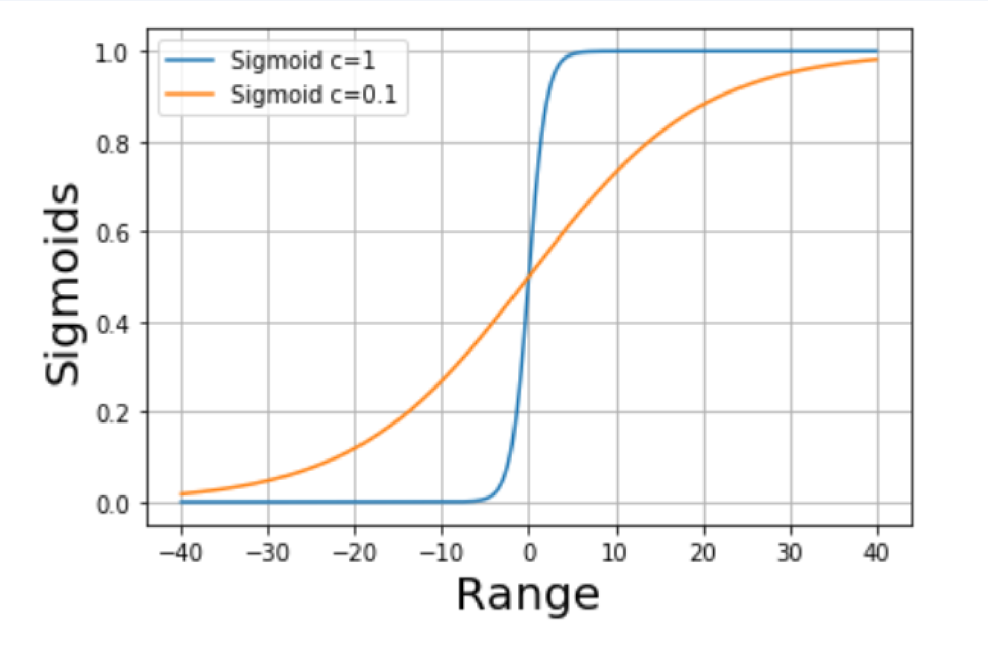
Estimate for Probability in Elo's Ranking

We use x= r(A) – r(B) and compute $P_{A>B}$ *= sigmoid(x)*

Properties of function f(x) that takes x and yields a probability:

- f(0) = ½
- f(infty) = 1
- f(-infty) = 0

$$f(x) = \frac{1}{1+e^{-cx}}$$



Statistics: Distributions

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.

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

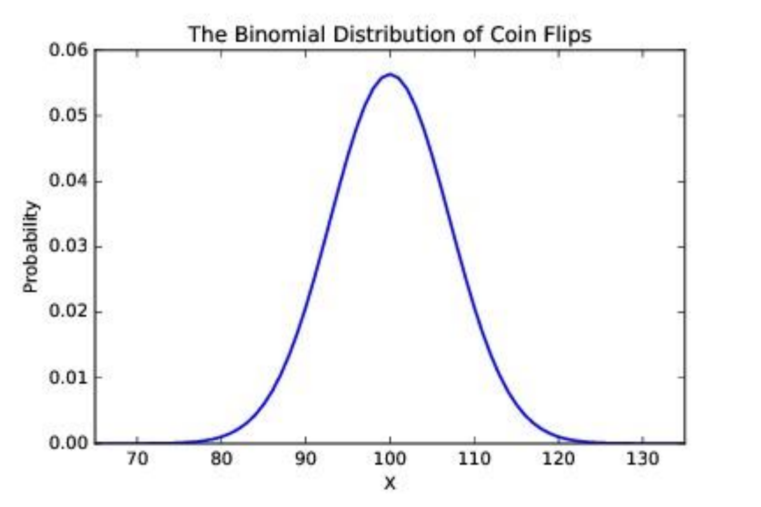
Applications:

- Number of games a team will win in a championship
- Number of defective products in a manufacturing unit
- Allocating resources on a web-server

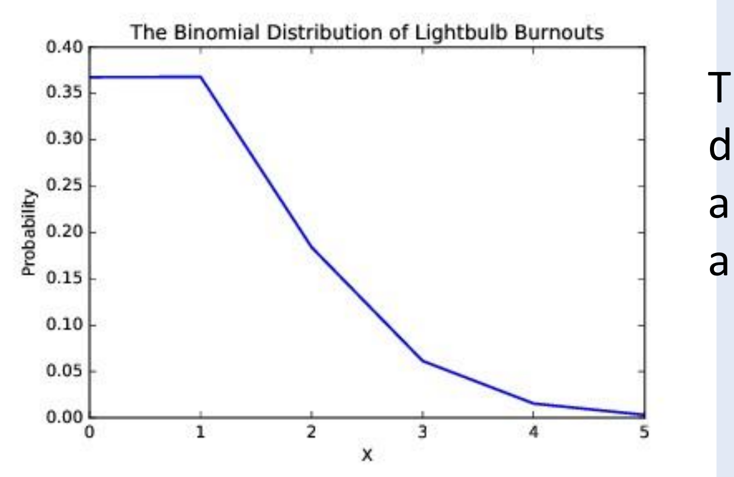
```
scipy.special.binom(100,i)*(p1**i)*(q1**(100-i))
np.random.binomial(10, 0.5, size=10)
```

Properties of Binomial Distributions

Discrete, but bell (or half-bell) shaped depending on ‘p’



Coin flips: p=0.5   n=200



Lightbulb burnouts: p=0.001 n=1000

The distribution is a function of n and p.

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

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

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

Geometric, Negative Binomial, Hypergeometric

- **Geometric:** Constructed from independent Bernoulli trials. X is the total number of trials up to and including the first success.
  - $P(k) = (1-p)^{k-1}p$ $\qquad$ $k=1,2,3\ldots$

- **Negative Binomial:** Trials until *r* successes are achieved
- **Hypergeometric Distribution:** Suppose an urn contains *n* balls, *r* are black and *n-r* are white. Let X denote the number of black balls taken out when *m* balls are withdrawn without replacement.

The Poisson Distribution

The Poisson distribution measures the number of times an event occurs in a given interval, such as number of telephone calls per minute of number of errors per page in a document

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

Instead of event probability *p*, the distribution is parameterized by mean μ, but this is equivalent because

$$\mu = \sum_{k=0}^{\infty} k \cdot Pr(k)$$

Assumptions behind Poisson

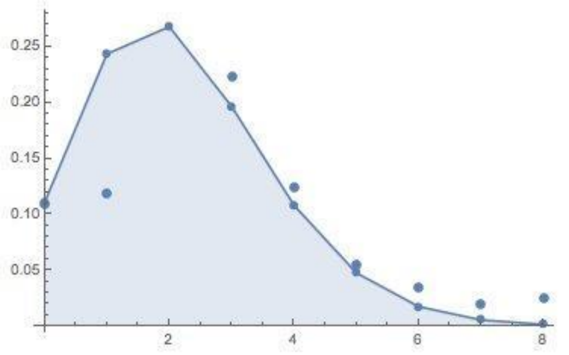
- What happens in one sub-interval is independent of what happens in any other sub-interval
- The probability of an event is the same in each sub-interval
- Events don’t happen simultaneously

- Poisson frequency function can be used to approximate binomial probabilities for large n and small p; with np= μ

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:



Exponential Distribution

- Density function: f(x) = $\begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$

- Used to model lifetimes or waiting times
- Memoryless
- Related to Poisson process

The Normal Distribution

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

Completely parameterized by mean and standard deviation.

If X~N(μ,σ^2) then (X- μ)/σ ~ N(0,1)

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

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

```
np.random.randn(10000)
```

Properties of the Normal Distribution

- It is a generalization of the binomial distribution where $n \rightarrow \infty$
- 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 independent normally distributed variables is normal.

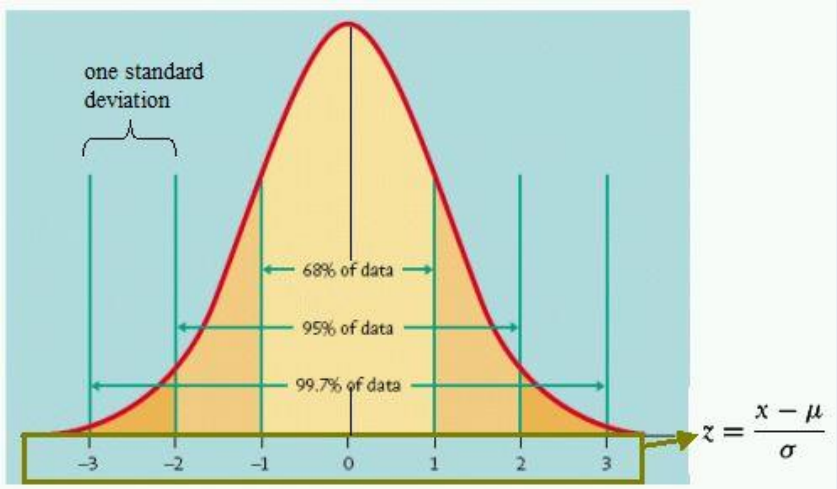
Question: Is it justified to use a normal distribution for +ve random variables?

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.

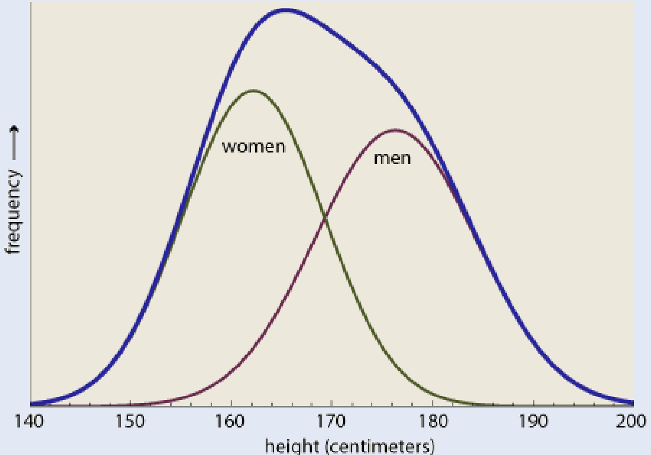


What’s not Normal?

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



χ^2

- If Z is a standard normal random variable, the distribution of $U=Z^2$ is called the chi-squared distribution with 1 degree of freedom.
- If U_1 ,U_2 …. U_n are n independent χ^2 random variables with 1 degree of freedom, the distribution of $V= U_1 +U_2 +\ldots + U_n$ is called the chi-squared distribution with n degrees of freedom and denoted by χ_n^2

t and F distributions

- If Z ~ N(0,1) and U~ χ_n^2 and Z and U are independent, then the distribution of $\frac{Z}{\sqrt{\frac{U}{n}}}$ is called the t distribution with n degrees of freedom. Also called student’s ‘t’ distribution
- Let U and V be independent chi-squared random variables with m and n degrees of freedom respectively, The distribution of W= $\frac{U/m}{V/n}$ is called the F distribution with m and n degrees of freedom and is denoted by $F_{m,n}$

Gamma Density Function

- g(t) = $\frac{\lambda^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}$, t>=0
- For t<0, g(t) =0
- α is the shape parameter and λ is the scale parameter
- $\Gamma(x)$=$\int_0^{\infty} u^{x-1} e^{-u}$du, x>0

- **Exercise:** Plot Gamma for various values of α and λ
- For α=1, it corresponds to the exponential density function

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.

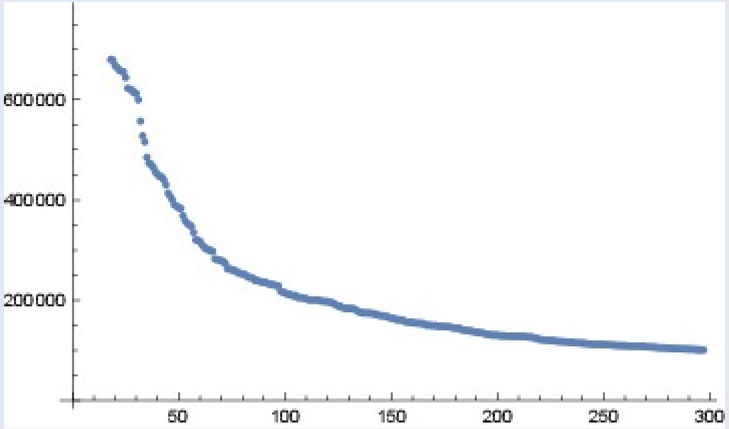
Exercise: Is it the same for India?

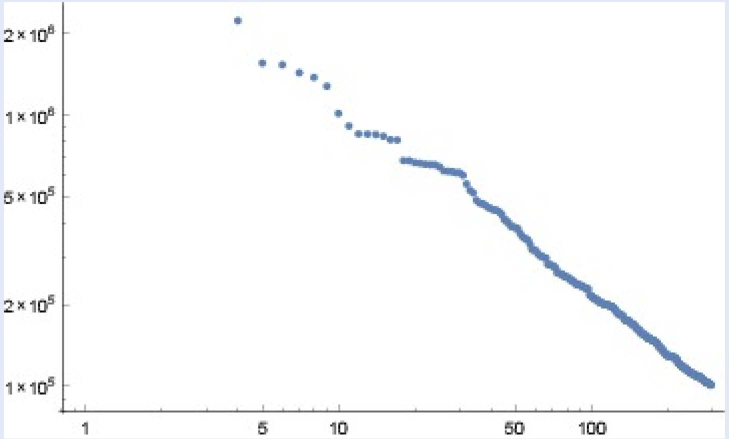
url = "https://www.census2011.co.in/city.php"

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,

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

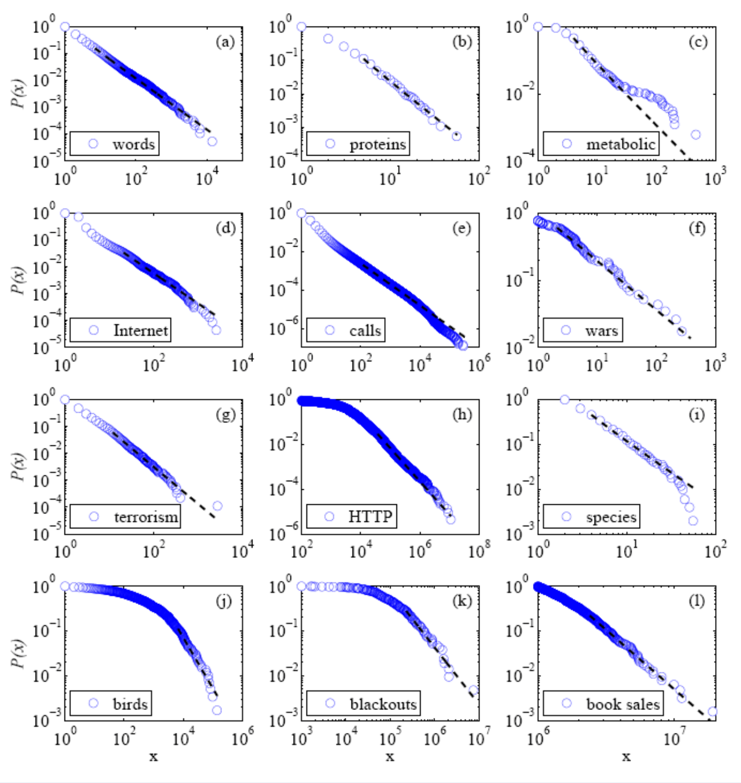
The constant *c* is unimportant: for a given *a* this constant *c* ensures the probability sums to 1.

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

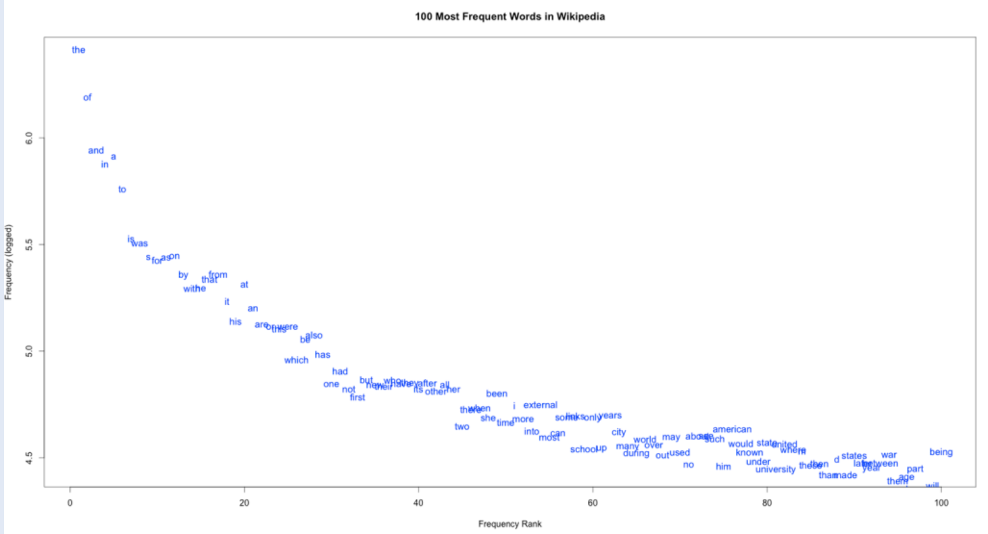
Power laws show as straight lines on log value, log frequency plots.



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.*

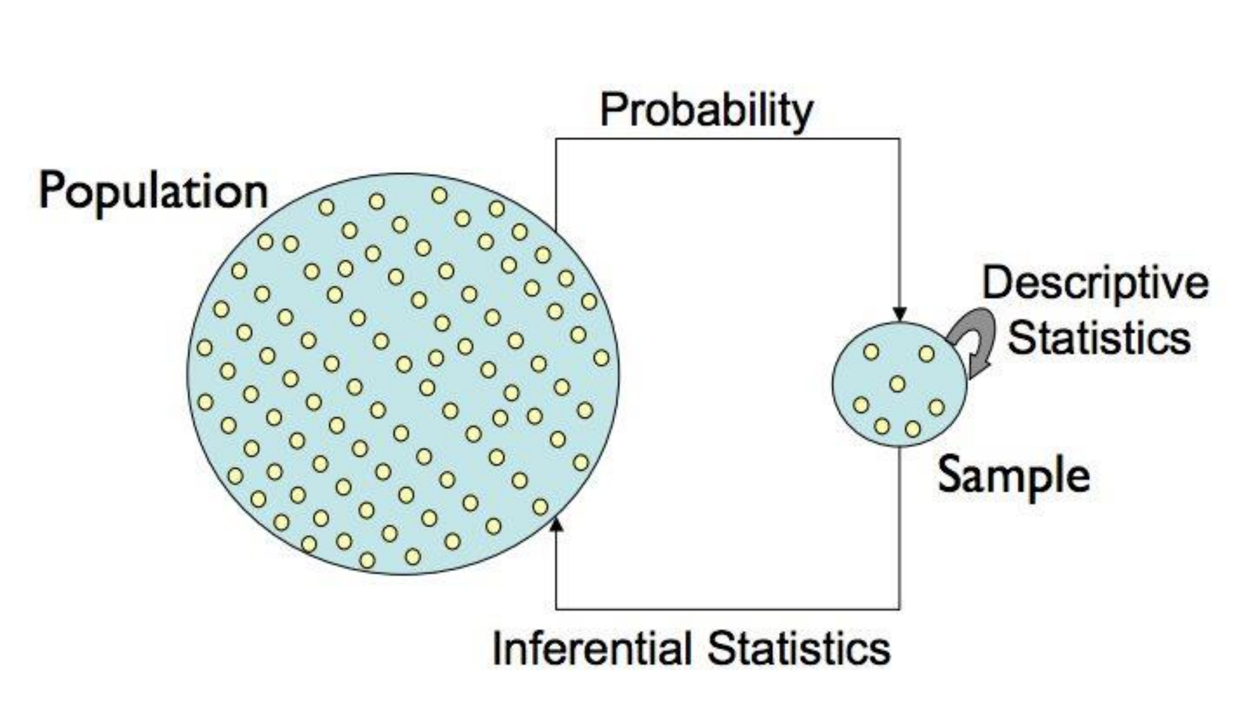


Properties of Power Laws

- 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.

Sampling from a distribution

The Central Dogma of Statistics

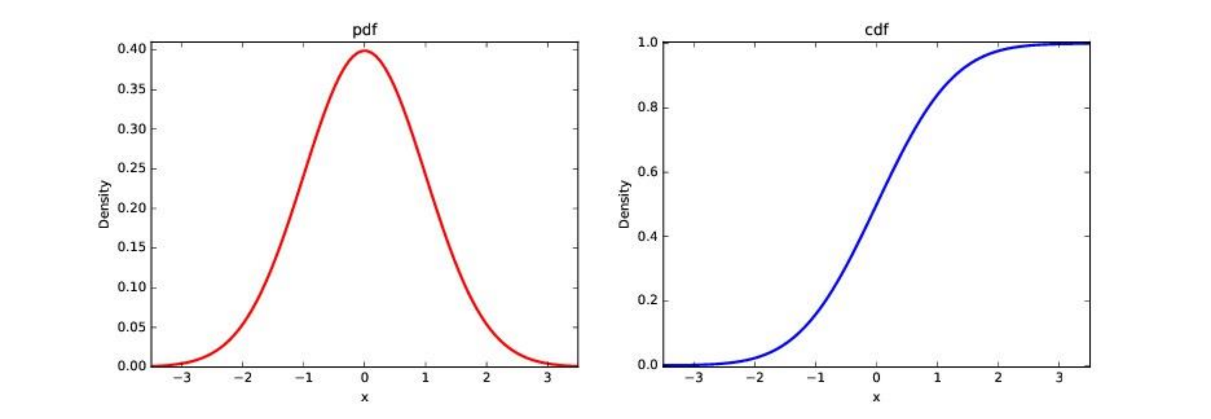


Source: www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture2.pdf

Sampling in One Dimension

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:

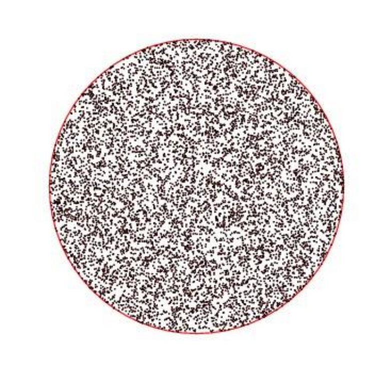


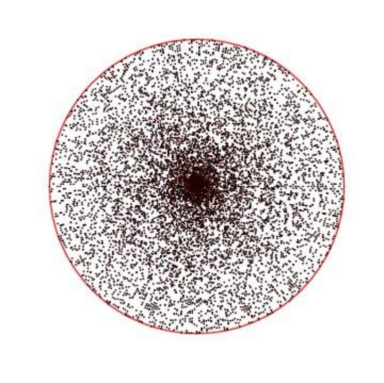
Sampling from Distributions

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.

Correlation Analysis

Definitions

- Suppose we are given two variables X and Y, represented by a sample of n points of the form (x_i; y_i), for 1<= i <= n.
- X and Y are **correlated** when the value of X has some predictive power on the value of Y.
- Correlation coefficient **r(X; Y)**
  - Measures the degree to which Y is a function of X, and vice versa.
  - Its value ranges from -1 to 1
  - 1 means fully correlated and 0 => n relation, i.e. independence
  - -ve => variables are anti-correlated, if X goes up, Y goes down.

The Pearson Correlation Coefficient

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

- The elements of numerator are +ve when samples of both X and Y are greater than the mean or smaller at the same time. This would mean that they are positively correlated.
- If X is usually greater than the mean when Y is less than its mean, i.e. the signs are –ve, this would mean they are –vely correlated.
- Works well for linear predictors

The Spearman Rank Correlation Coefficient

- Count the number of pairs of input points which are out of order.
- Suppose the data set contains points $(x1; y1)$ and $(x2; y2)$ where $x1 < x2$ and $y1 < y2$.
- This is a vote that the values are positively correlated, whereas the vote would be for a negative correlation if $y2 < y1$.
- Summing up over all pairs of points and normalizing properly gives us Spearman rank correlation.
- This method is much more robust against outliers

The Spearman Rank Correlation Coefficient

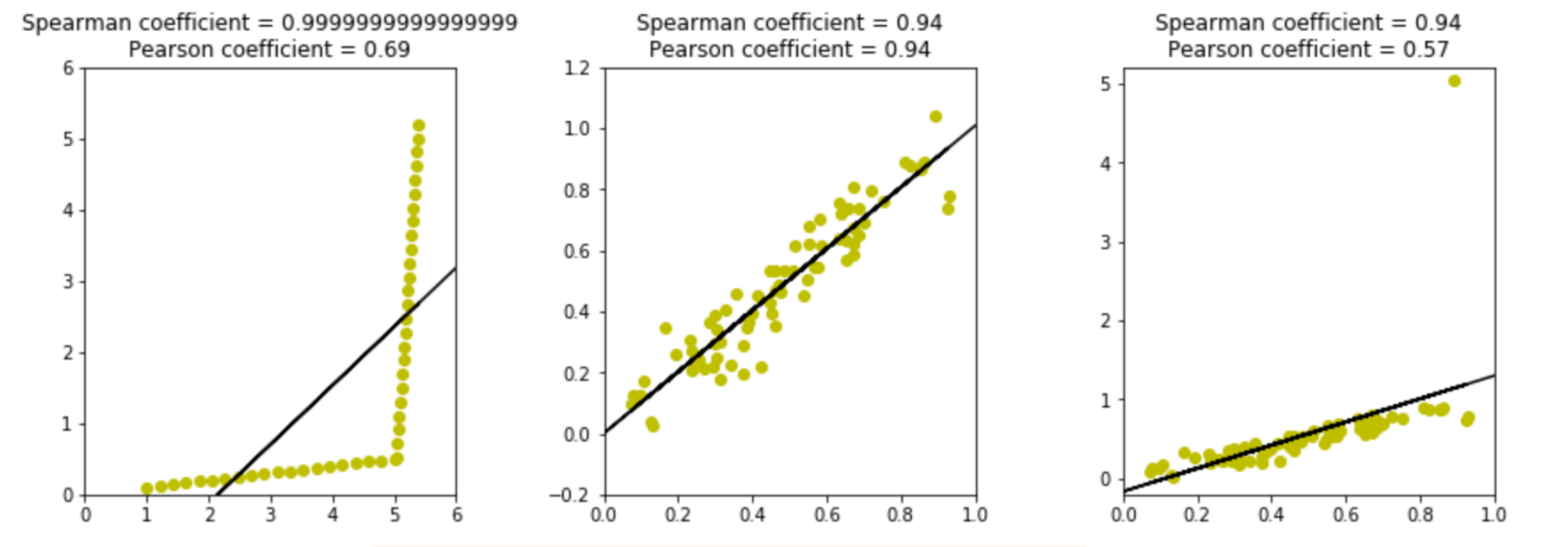
- Let rank(x_i) be the rank position of x_i in sorted order among all x_i, so the rank of the smallest value is 1 and the largest value n.
- Then spearman rank correlation coefficient is,

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where $d_i = rank(x_i) - rank(y_i)$.

Exercise: Prove that -1 < ρ <1 for any value of n

Comparison



Spearman coefficient = 0.9999999999999999
Pearson coefficient = 0.69

Spearman coefficient = 0.94
Pearson coefficient = 0.94

Spearman coefficient = 0.94
Pearson coefficient = 0.57

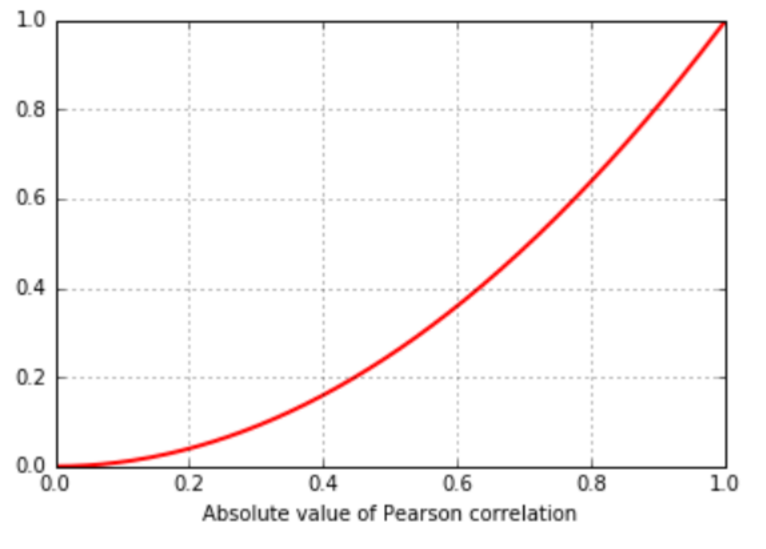
```
rho, pval = stats.spearmanr(xval, yval)
pearson = stats.pearsonr(xval, yval)
```

True or False?

- Taller people are more likely to remain lean
- Medicine causes illness
- Standardized tests like GRE, Board exams predict the performance of students in college/university
- Financial status affects health
- Smoking causes cancer
- Active Police in an area increases crime
- Wealth decreases family size

Strength of Correlation

- R^2 : The strength of the correlation is defined as the square of the sample correlation coefficient. It estimates the fraction of variance in Y explained by X in a **simple linear regression**.



Statistical significance

- The statistical significance of a correlation depends upon its sample size *n* as well as *r*.
- A correlation of n points is significant if there is an $\alpha \leq \frac{1}{20} = 0.05$ chance that we would observe a correlation as strong as *r* in any random set of *n* points.

Hypothesis Testing

Motivation

- How do we evaluate a theory or hypothesis on the basis of data?
- Hypothesis testing is a methodology developed in the field of statistics.

As an example, 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.

A/B Testing Problem

- Evaluate whether a product change makes a difference in performance.
- Suppose you show one group of users version A, and another group version B. {Usually take a 50/50 random split}
- Measure system performance, eg:
  - Number of times they click on ads
  - Number of stars they give it when asked about the experience
- The t-test measures whether the observed difference between the two groups is significant

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?

When working with big-data models with lot of features which have weak correlation, any single feature may explain only small effects, but may be together they can have a strong predictive power.

Steps

- Formulate the hypothesis in terms of a **null hypothesis**, H_o, which represents the currently accepted state of knowledge
- An **alternative hypothesis**, H_A , which represent a new claim that challenges the current state of knowledge.
- The null hypothesis and the alternative hypothesis must be mutually exclusive and complementary, so that one and only one of the hypotheses are true.

Steps

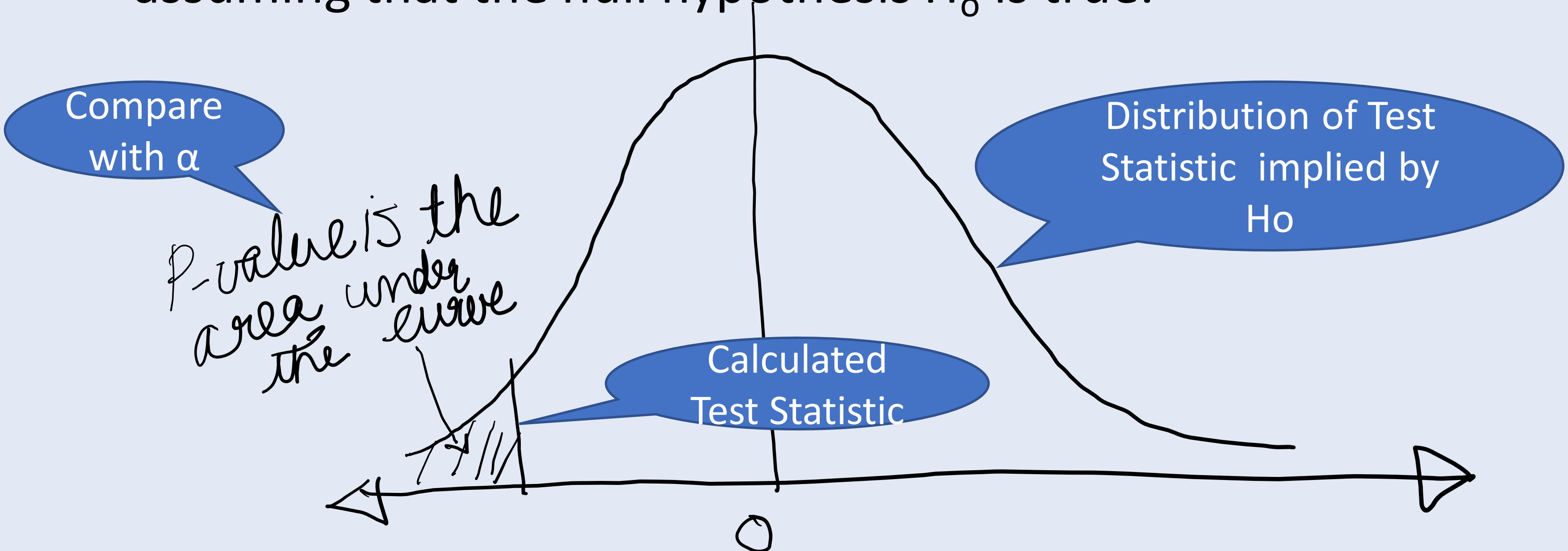
- Once H_o and H_A are defined, data that support the test must be collected, for example, through measurements, observations, or a survey.
- The next step is to find a test statistic that can be computed from the data (example sample mean), and whose probability distribution function can be found under the null hypothesis.
- Next, we can evaluate the data by computing the probability (the ***p-value***) of obtaining the observed value of the test statistic (or a more extreme one) using the distribution function that is implied by the null hypothesis.

P-value

- If the *p*-value is smaller than a predetermined threshold, known as the significance level, and denoted by α (typically 5% or 1%), we can conclude that the observed data is unlikely to have been described by the distribution corresponding to the null hypothesis.
- In that case, we can therefore reject the null hypothesis in favor of the alternative hypothesis.

P-value

- **Definition:** Probability of observing a test statistic or a value more extreme than the ones observed in your data, assuming that the null hypothesis H_o is true.

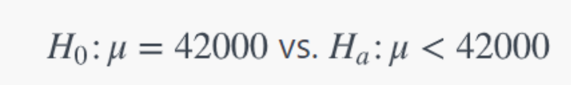

Compare with α
P-value is the area under the curve
Calculated Test Statistic
Distribution of Test Statistic implied by Ho
0

Errors in Hypothesis Testing

- Statistical hypothesis testing is a probabilistic method, which means that we cannot be certain in the decision to reject or not to reject the null hypothesis.
- There can be two types of error:
  - Type 1: Reject the null hypothesis when it is true (Prob of error α)
  - Type 2: Fail to reject null hypothesis when it should be rejected
- By choosing the required significance level we can balance the trade-off between these two types of error

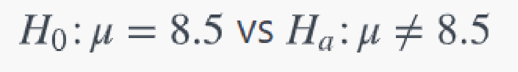
Construct Null and Alternate Hypothesis

- A consumer test agency wants to see the whether the mean lifetime of a brand of tires is less than 42,000 miles. The tire manufacturer advertises that the average lifetime is at least 42,000 miles.



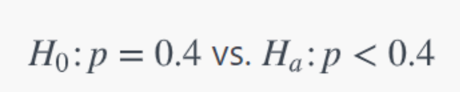
$H_0 : \mu = 42000$ vs. $H_a : \mu < 42000$

- The length of a certain lumber from a national home building store is supposed to be 8.5 feet. A builder wants to check whether the shipment of lumber she receives has a mean length different from 8.5 feet.



$H_0 : \mu = 8.5$ vs $H_a : \mu \neq 8.5$

- A political news company believes the national approval rating for the current president has fallen below 40%.



$H_0 : p = 0.4$ vs. $H_a : p < 0.4$

https://online.stat.psu.edu/stat500/lesson/6a/6a.4/6a.4.1

Comparing Population Means

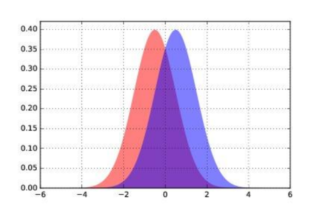
The T-test evaluates whether the population means of two samples are different.

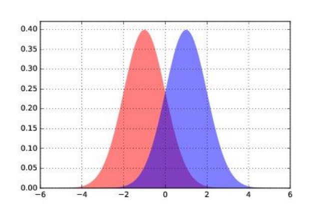
Sample the IQs of 20 men and 20 women. Is one group smarter on average?

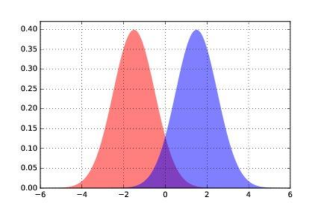
Certainly the sample means will differ, but is this difference significant?

Differences in Distributions

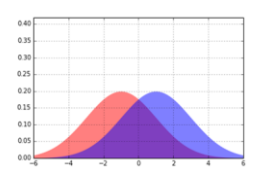
It becomes easier to distinguish two distributions as the means move apart...

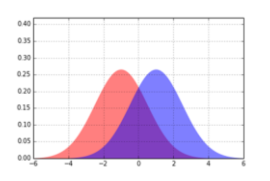


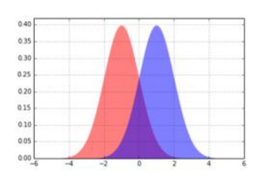




... or the variance decreases:



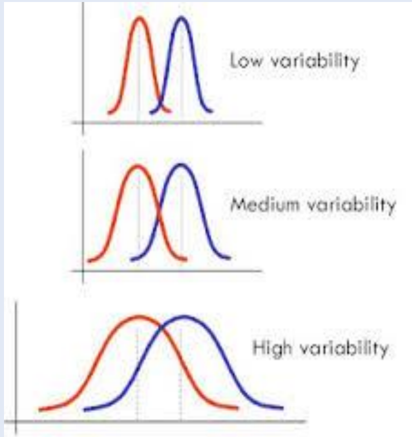




The T-Test

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.

The table entry specifies the value v that the t-statistic t must exceed.

If t > v, then the observation is significant to the α level.

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.
- If the sample is too far from the mean, it must be from a different distribution…
- Bayes theorem converts Pr(data|distribution) to Pr(distribution|data)

The Kolmogorov-Smirnov Test

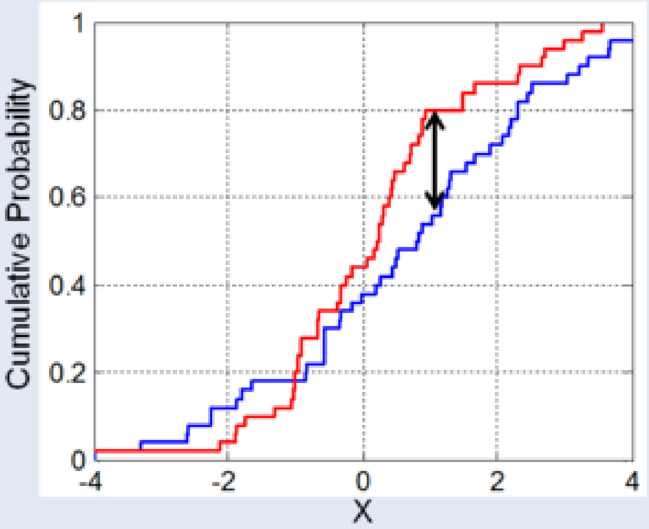
This test measures whether two samples are drawn from same distribution by the maximum difference in their cdf.

The distributions differ if:

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

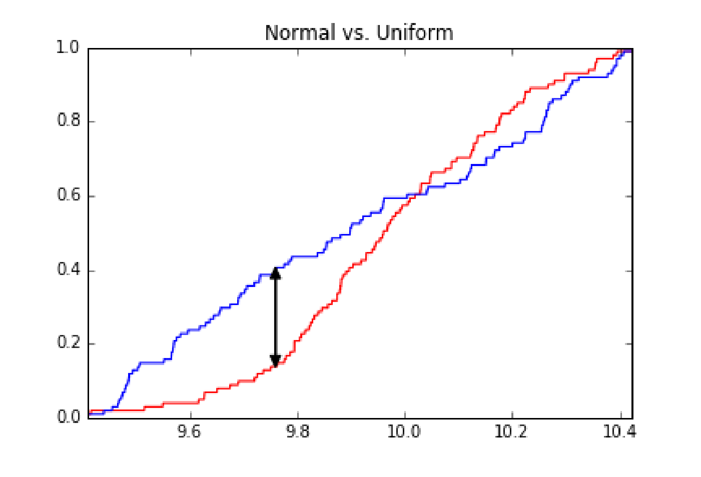
at a significance of alpha where, c(α) is looked up in a table.

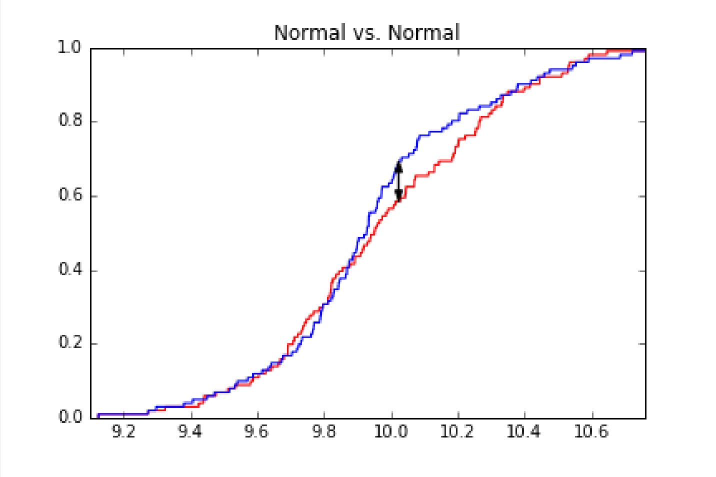
$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|,$$



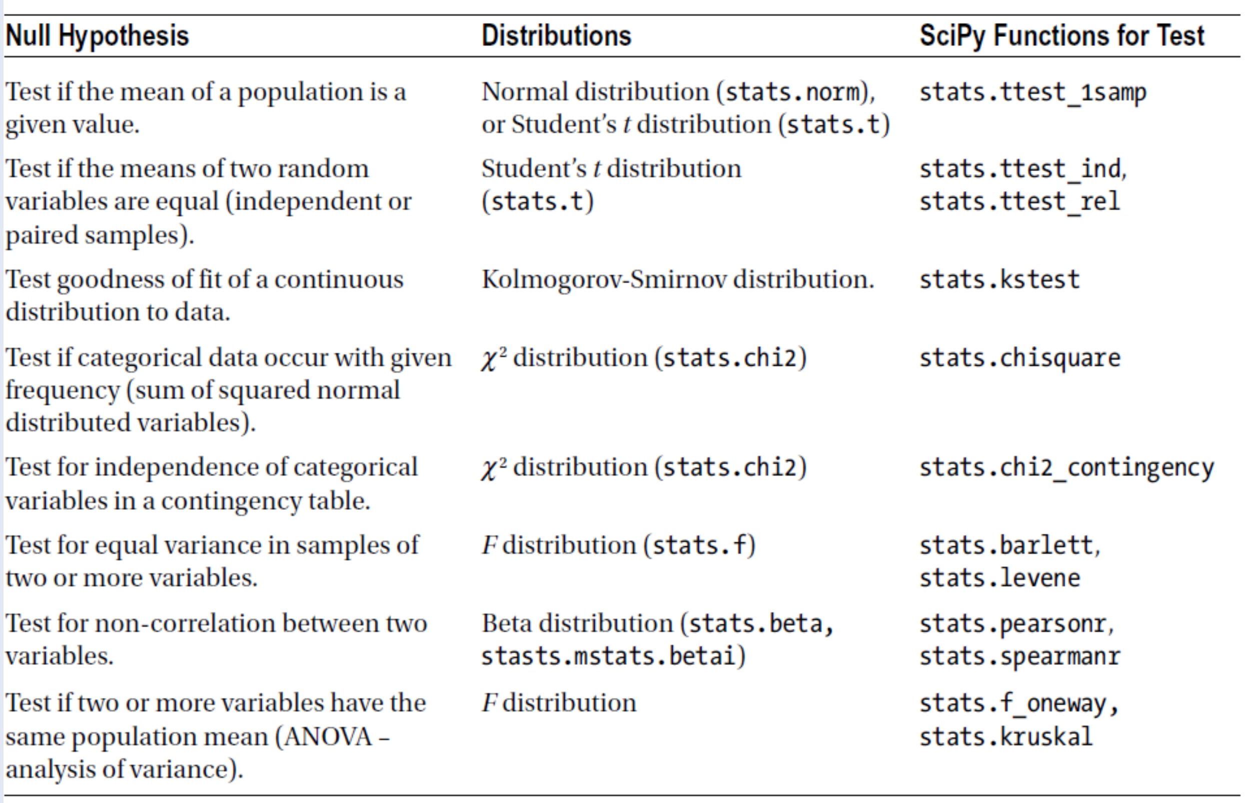
Normality Testing

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





Different Hypothesis Tests



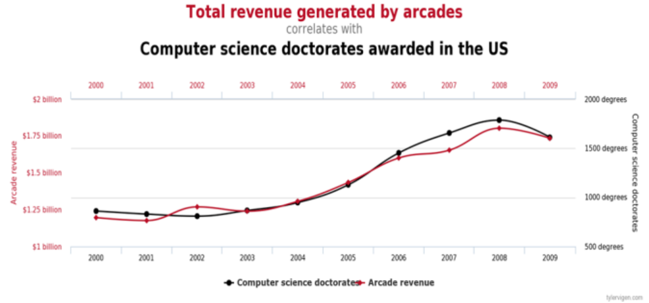
Null Hypothesis	Distributions	SciPy Functions for Test
Test if the mean of a population is a given value.	Normal distribution (`stats.norm`), or Student's *t* distribution (`stats.t`)	`stats.ttest_1samp`
Test if the means of two random variables are equal (independent or paired samples).	Student's *t* distribution (`stats.t`)	`stats.ttest_ind`, `stats.ttest_rel`
Test goodness of fit of a continuous distribution to data.	Kolmogorov-Smirnov distribution.	`stats.kstest`
Test if categorical data occur with given frequency (sum of squared normal distributed variables).	χ^2 distribution (`stats.chi2`)	`stats.chisquare`
Test for independence of categorical variables in a contingency table.	χ^2 distribution (`stats.chi2`)	`stats.chi2_contingency`
Test for equal variance in samples of two or more variables.	*F* distribution (`stats.f`)	`stats.barlett`, `stats.levene`
Test for non-correlation between two variables.	Beta distribution (`stats.beta`, `stasts.mstats.betai`)	`stats.pearsonr`, `stats.spearmanr`
Test if two or more variables have the same population mean (ANOVA – analysis of variance).	*F* distribution	`stats.f_oneway`, `stats.kruskal`

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*.



Significance and Effect Size

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

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

We informally categorize a medium-level effect size as visible to the naked eye by a careful observer. On this scale, large effects pop out, and small effects are not completely trivial

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

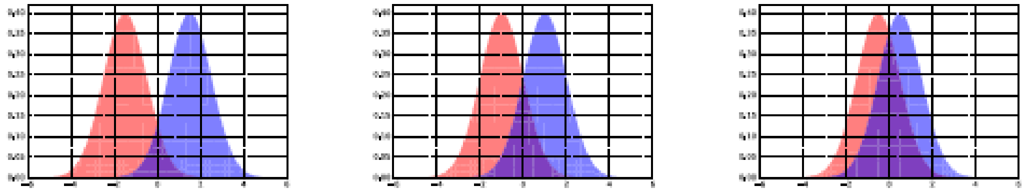


Figure 5.11: Pairs of normal distributions with the same variance, but decreasing difference in their means from left to right. As the means get closer, the greater overlap between the distributions makes it harder to tell them apart.

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.

They also allow us to construct non-traditional test-statistics.

Permutation Tests

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

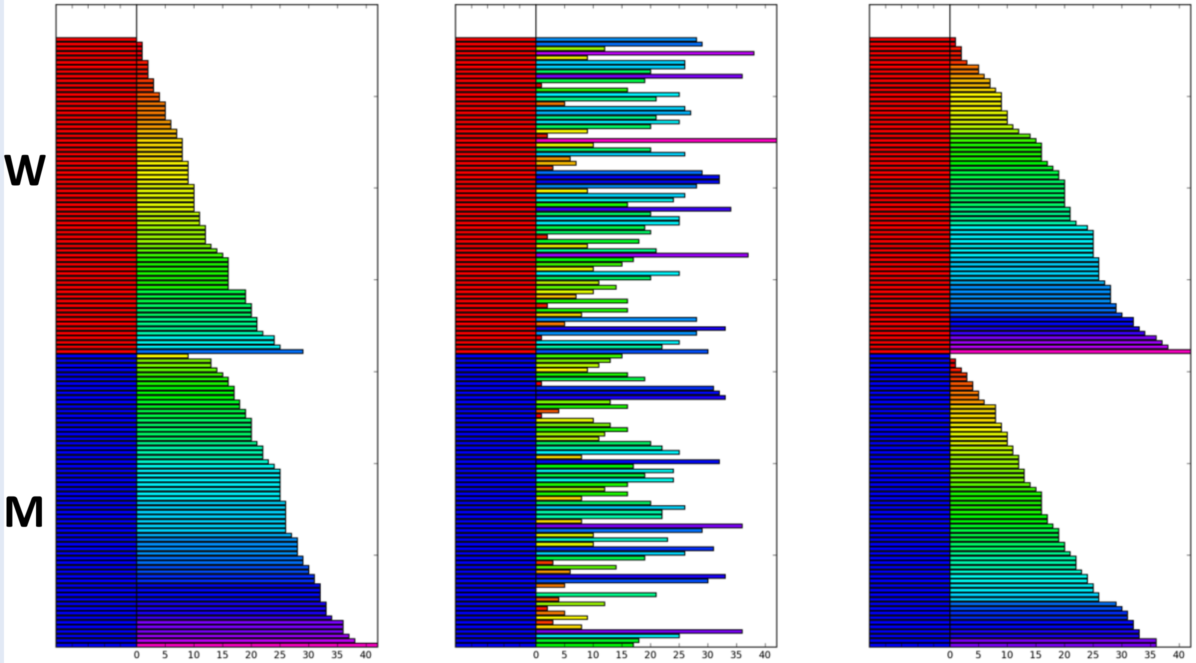
The ranking of the real test statistic among the shuffled test statistics gives a p-value.

You need a 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

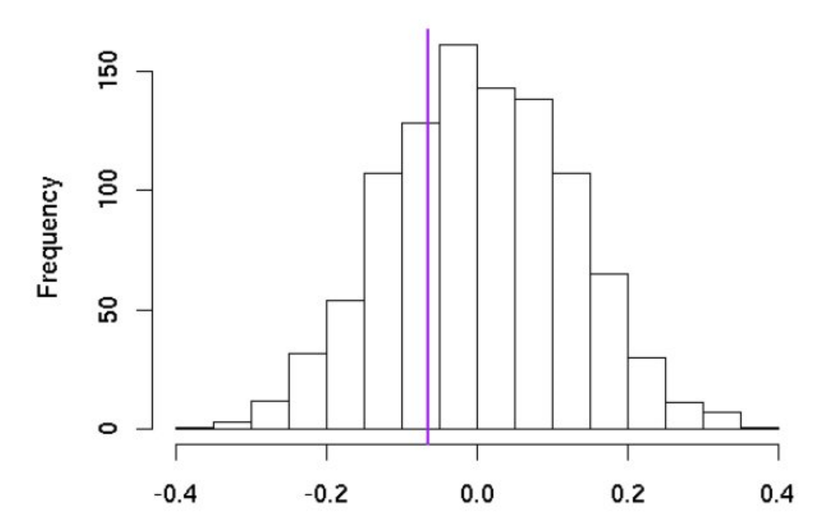
The random permutation (center/right) shows less height difference by gender than the original data (left).

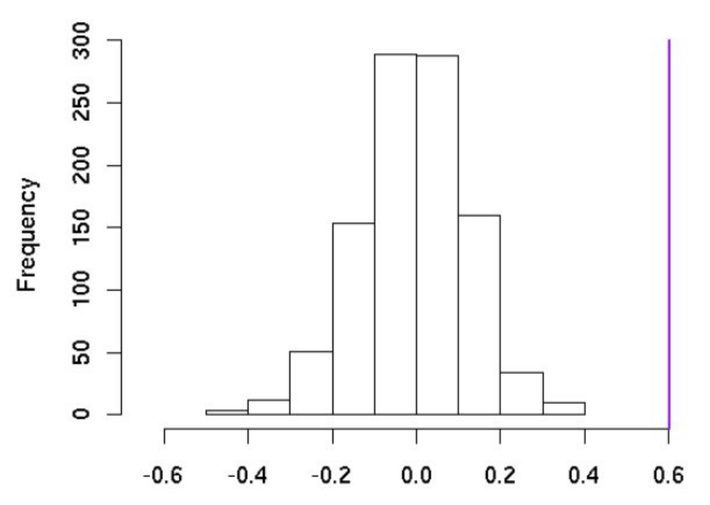


Significance of a Permutation Test

The rank of the real data among the random permutations determines significance. If it is towards the right, it is significant.

p-value = rank of observed test-statistic/(Total Permutations)





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!