大数据分析实践

Statistical Methods

Qiong Zeng (曾琼)

qiong.zn@sdu.edu.cn

Research *is creative and systematic work undertaken to increase the stock of knowledge.*


Visual C#
C++
Curl
Nuva
.NET
Ruby
MATLAB
Visual Basic
DMDScript
Delphi
SQL
Jess
J#
Golang
C++/CLI
Scala
Objective-C
Smalltalk
bash
Prolog
ALGOL
BASIC
ActionScript
B
Haskell
LISP
PowerShell
HTML
JAVA
F#
Self
Perl
Erlang
C#
JavaScript
JScript
IronPython
Pascal
Small Basic
XML
Visual Basic
Swift
Groovy
Python
VBScript
PHP
Lazarus
Kotlin
C
Borland C++
FORTRAN
Scheme
Lua
COBOL

Course Outline

6 Personal assignments
6 Group assignments

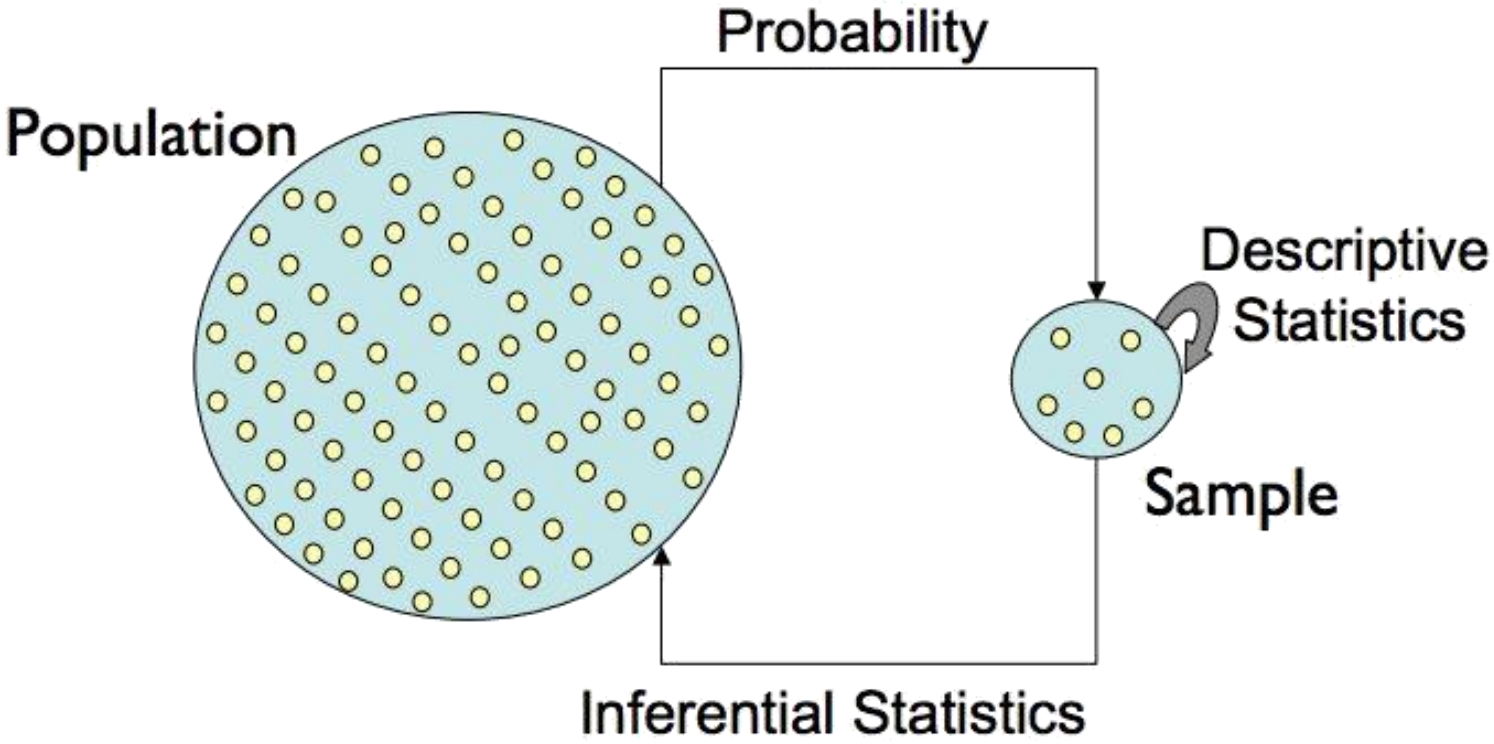
上课日期	授课内容	实验内容	周次
20240905	课程入门、大数据探索式分析	/	第一周
20240912	课程实践项目介绍、项目组队测试、项目经验谈	项目成员集结	第二周
20240919	科研实践入门、数据采样与降维	项目管理工具制定项目计划、**Pandas数据采样实践**	第三周
2024926	数据质量管理	Pandas数据质量实践	第四周
20241003	/	/	第五周
20241010	众包与电子表格	电子表格实践	第六周
20241017	可视化设计	可视化设计实践	第七周
20241024	**统计分析方法与工具**	**统计方法实践**	**第八周**
20241031	中期汇报（论文+项目进展）1	中期进展报告	第九周
20241107	中期汇报（论文+项目进展）2	BERT实践环境配置	第十周
20241114	机器学习方法与工具	BERT实践	第十一周
20241121	人机交互方法与工具	Canis/Cast/Libra实践	第十二周
20241128	普适计算	手机移动数据采集与分析	第十三周
20241205	大规模数据分析系统	SPARK实践	第十四周
20241212	如何撰写项目论文	大项目收尾	第十五周
20241219	项目结题报告1	大项目验收	第十六周

学习目标

- 可复述基本统计分布的定义，能够描述不同统计分布的区别，能后辨别不同数据集上所适合的分布类型
- 阐述统计显著性检验定义以及基本方法
- 了解多元分析

Statistical Distributions

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（二项分布）
- The Normal Distribution（正态分布）
- The Poisson Distribution（泊松分布）
- The Power Law Distribution（幂律分布）

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

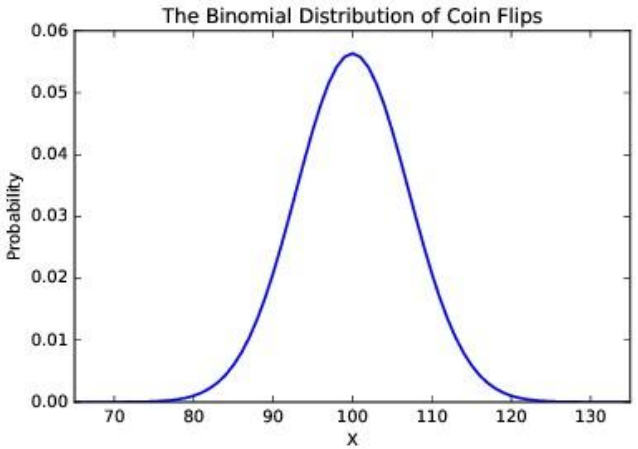
Experiments consist of *n identical, independent* trials which have two possible outcomes, with probabilities *p* and *(1-p)* like heads or tails.

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

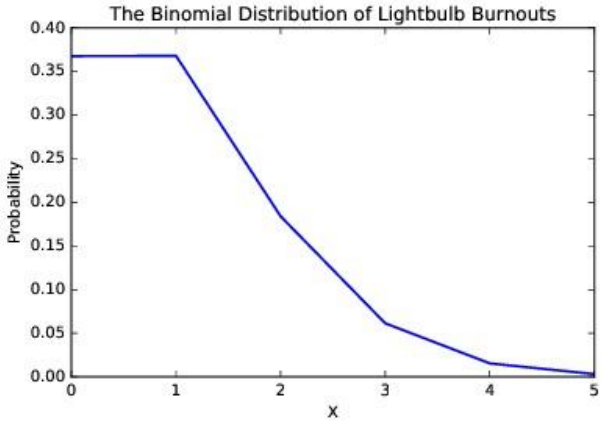
Properties of Binomial Distributions

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

$$n \to \infty$$

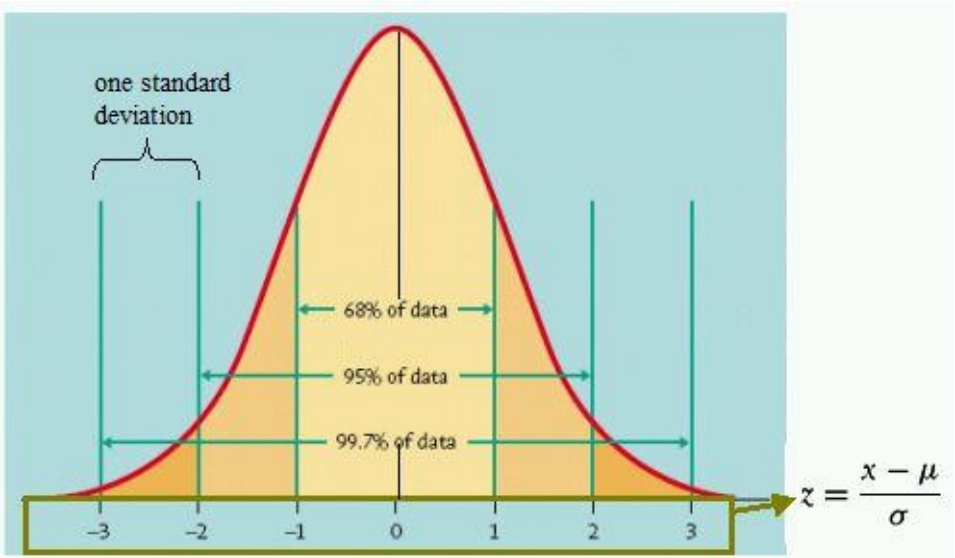
- Instead of *n* and *p*, the parameters are the mean *mu* and standard deviation *sigma*.
- It really **is** bell-shaped since *x* is continuous and goes infinitely in each direction.
- The sum of independent normally distributed variables is normal.

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.

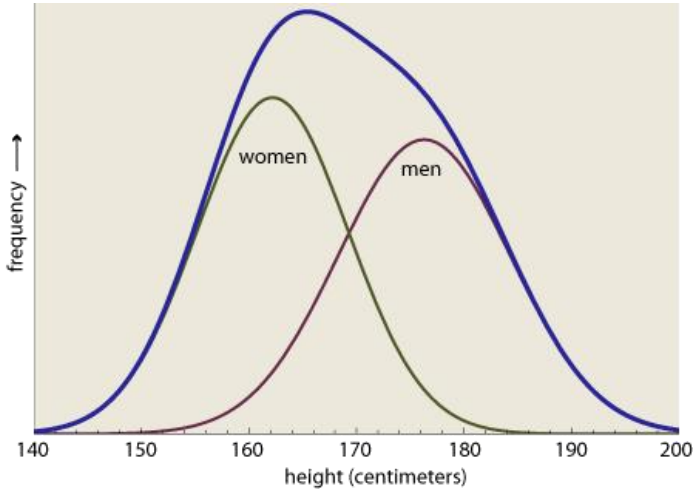


What’s not Normal?

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, 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.

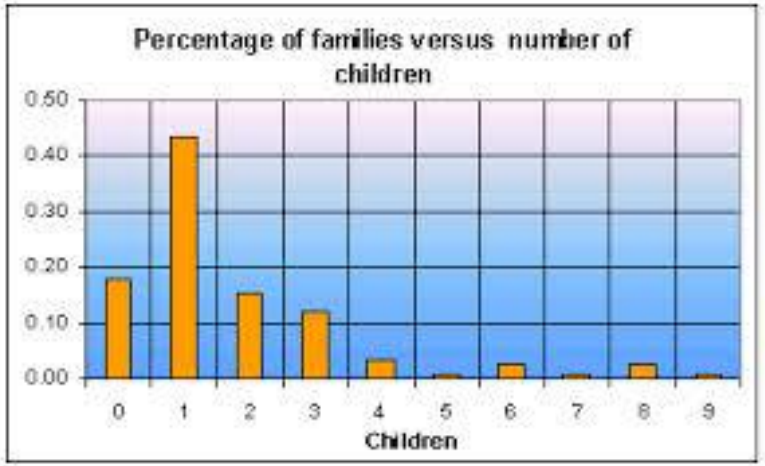
$$Pr(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

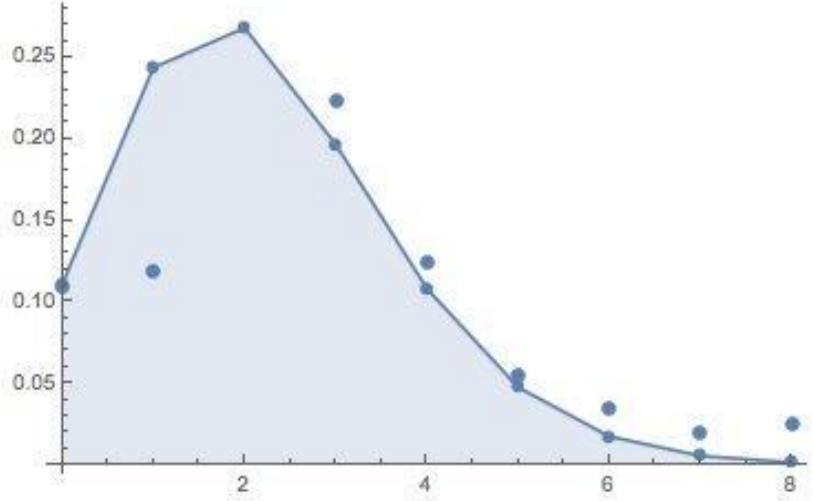
Instead of event probability *p*, the distribution is parameterized by mean *mu*

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.

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,

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

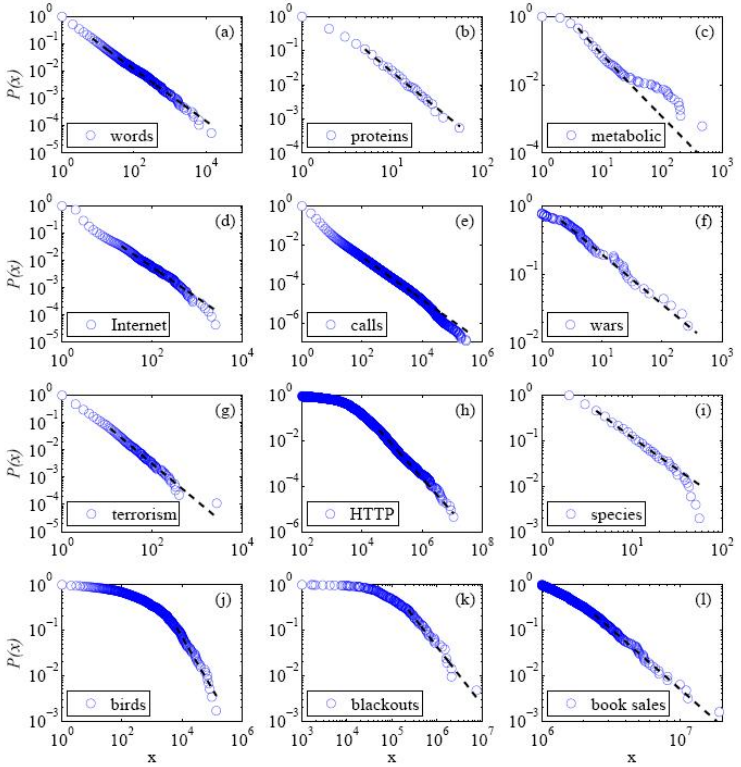
The constant *c* is unimportant: for a given *a* this constant *c* ensures the probability sums to 1.

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

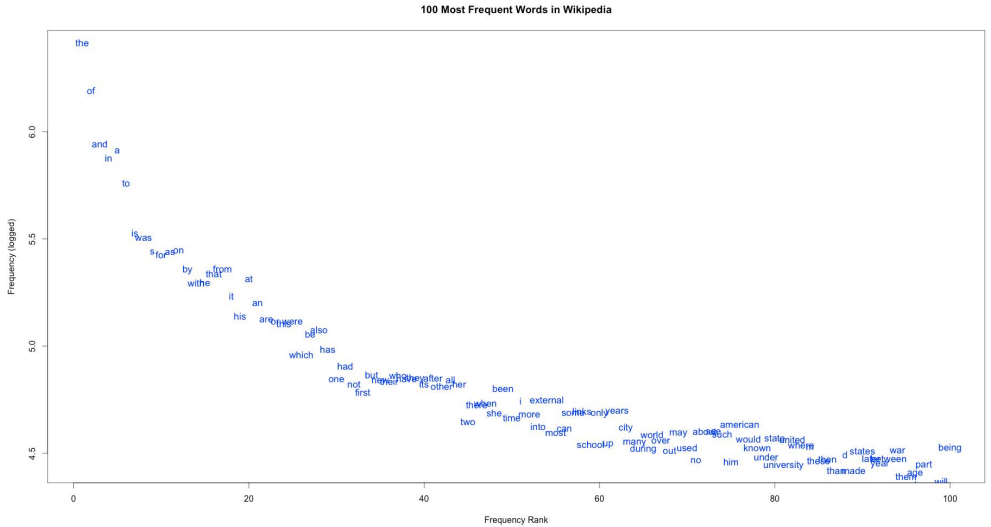
Power laws show as straight lines on log value, log frequency plots.



Word Frequencies and Zipf’ s Law

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

Zipf’s law is a power law for *α=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.

填空题 3分

设置

经典分布中哪个分布最适合描述以下现象，请说明原因：

（a）20岁程序员头发数量 [填空1]

（b）被闪电击中x次的人数 [填空2]

（c）早八课的出勤率 [填空3]

作答

Statistical Significance

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.

When ul?

Compu d

correlat

But wh

Sufficie s may

seem `` cts are

more su

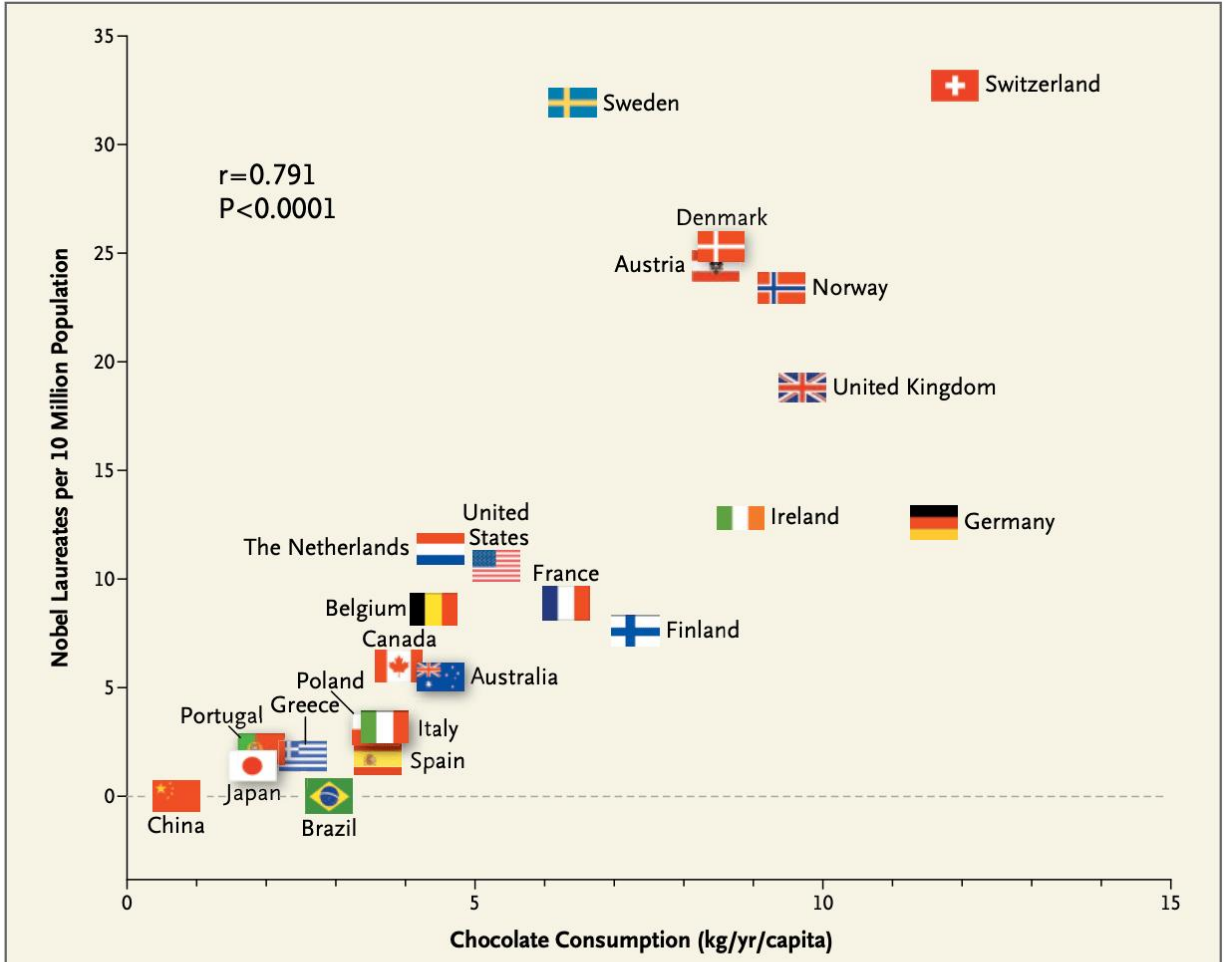


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

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.

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?

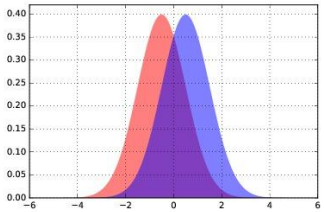
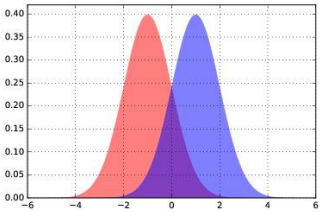
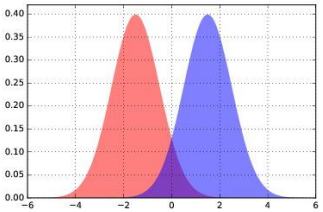
Measures of Effect Size（效应量）

- *Pearson correlation coefficient*（皮尔逊相关系数） small effects start at 0.2, medium effects at 0.5, large effects at 0.8
- *Percentage of overlap between distributions*（重叠百分比） small effects start at 53%, medium effects at 67%, large effects at 85%
- *Cohen's d* small >0.2, medium > 0.5, large > 0.8

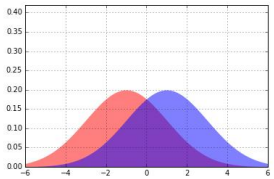
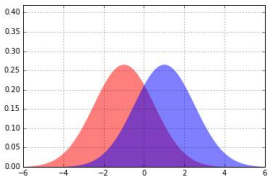
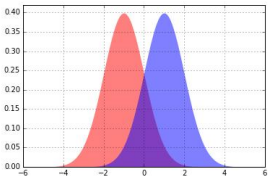
$$d = (|\mu - \mu')/\sigma)$$

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:

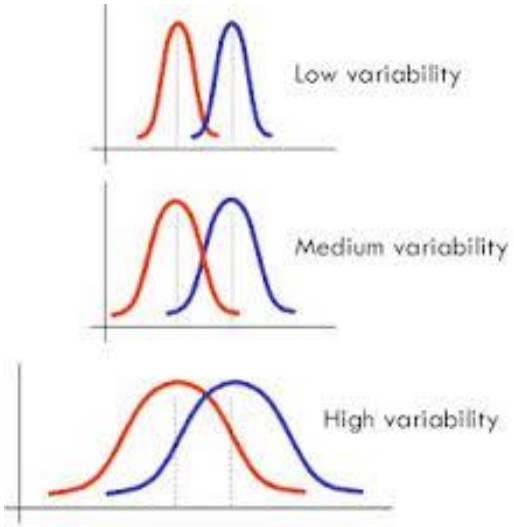
- The mean difference is relatively large
- The standard deviations are small enough
- The samples are large enough

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 Kolmogorov-Smirnov Test

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

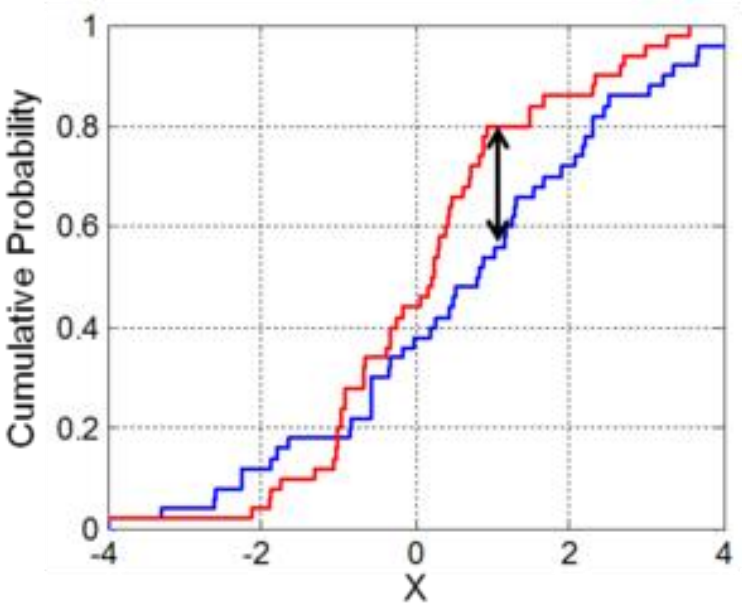
The distributions differ if:

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

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

and

at a significance of alpha.



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.

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.

Thank You

主观题 10分

Setting

性别	身高									
男生	185	175	170	180	187	175	186	192	169	171
女生	170	165	164	160	168	167	168	155	173	172

用**置换检验**确定男生的平均身高是否比女生的平均身高

正常使用主观题需2.0以上版本雨课堂

作答

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.

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 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).