Applied Data Analysis (CS401)

Lecture 5
Read the Stats Carefully
2018/10/18

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Robert West

dlab

Announcements

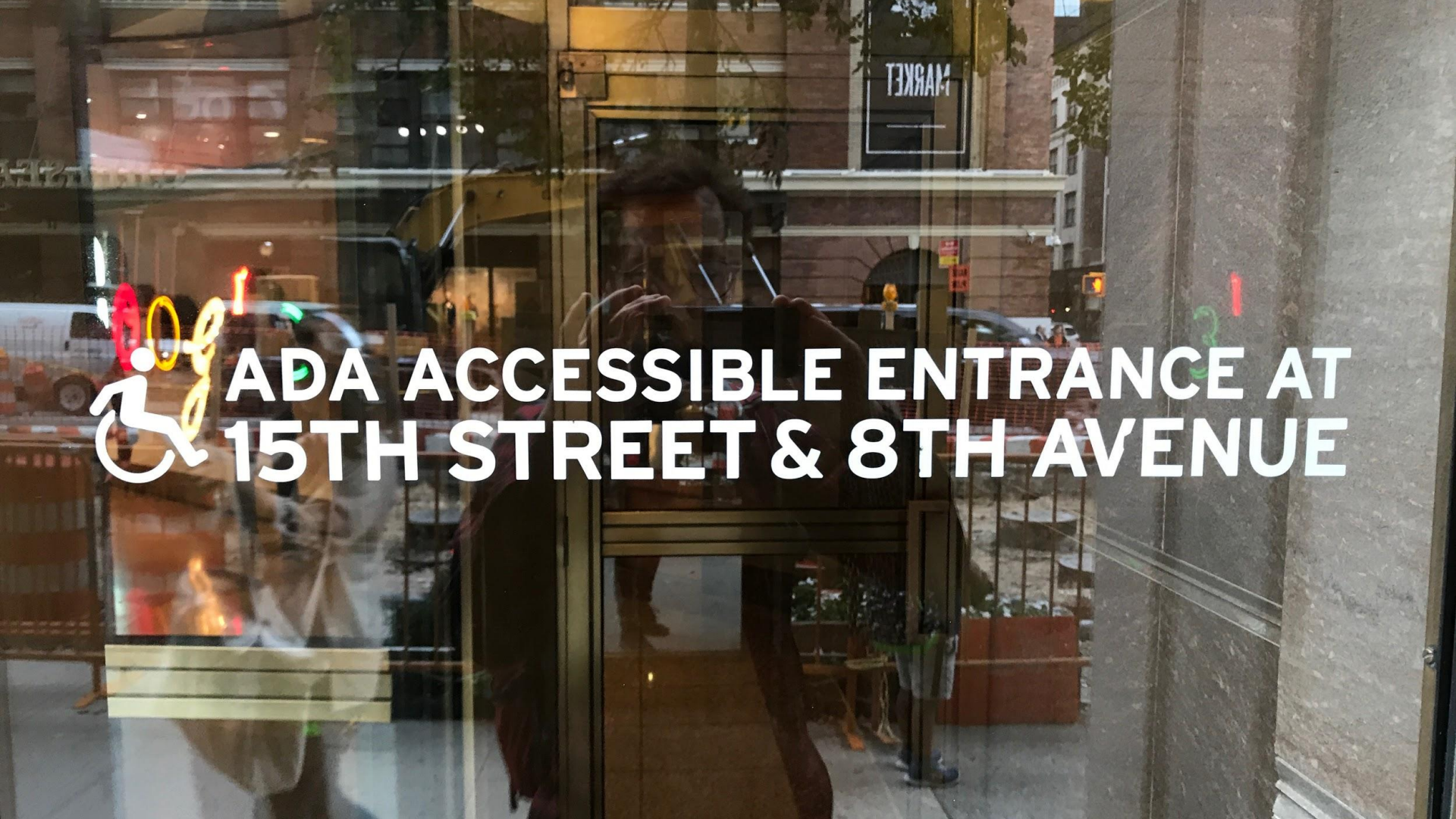
- Homework 1
  - Peer reviews due tomorrow (Fri) 23:59
  - Grades released Wed, Oct 24
- Homework 2
  - Due on Wed, Oct 24, 23:59
- Tomorrow’s lab session:
  - Project introduction
  - Cluster introduction
  - HW2 office hours
  - DataCluedo, installment 2

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2018-lec5-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- Where is Waldo?
- …


Google
INLAND TERMINAL NUMBER


MARKET
ADA ACCESSIBLE ENTRANCE AT
15TH STREET & 8TH AVENUE

Overview

- Descriptive statistics
- Sampling and uncertainty
- Relating two variables
- Hypothesis testing

ADA won't cover the basic of stats!

You know these things from prerequisite courses

But stats are a key ingredient of data analysis

Today: some highlights and common pitfalls

Descriptive statistics

Descriptive statistics

```
baseball.describe()
```

	year	stint	g	ab	r
count	100.00000	100.000000	100.000000	100.000000	100.00000
mean	2006.92000	1.130000	52.380000	136.540000	18.69000
std	0.27266	0.337998	48.031299	181.936853	27.77496
min	2006.00000	1.000000	1.000000	0.000000	0.00000
25%	2007.00000	1.000000	9.500000	2.000000	0.00000
50%	2007.00000	1.000000	33.000000	40.500000	2.00000
75%	2007.00000	1.000000	83.250000	243.750000	33.25000
max	2007.00000	2.000000	155.000000	586.000000	107.00000

Mean, variance, and normal distribution

The (arithmetic) **mean** of a set of values is just the average of the values.

Variance a measure of the width of a distribution. Specifically, the variance is the mean squared deviation of points from the mean:

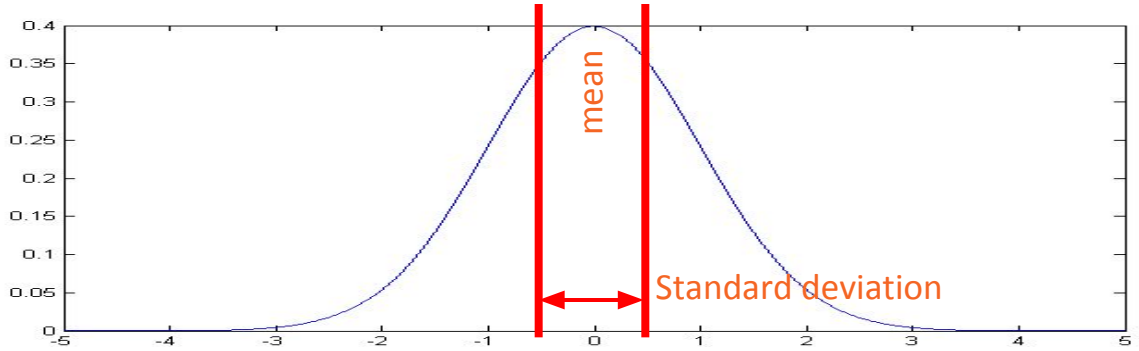
$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

The **standard deviation** (std) is the square root of variance.

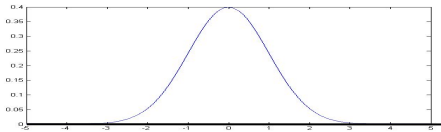
The normal distribution is completely characterized by mean and std.

```
baseball.describe()
```

	year	stint	g	ab	r
count	100.00000	100.000000	100.000000	100.000000	100.00000
mean	2006.92000	1.130000	52.380000	136.540000	18.69000
std	0.27266	0.337998	48.031299	181.936853	27.77496
min	2006.00000	1.000000	1.000000	0.000000	0.00000
25%	2007.00000	1.000000	9.500000	2.000000	0.00000
50%	2007.00000	1.000000	33.000000	40.500000	2.00000
75%	2007.00000	1.000000	83.250000	243.750000	33.25000
max	2007.00000	2.000000	155.000000	586.000000	107.00000



Robust statistics



A statistic is said to be robust if it is not sensitive to **outliers**

```
baseball.describe()
```

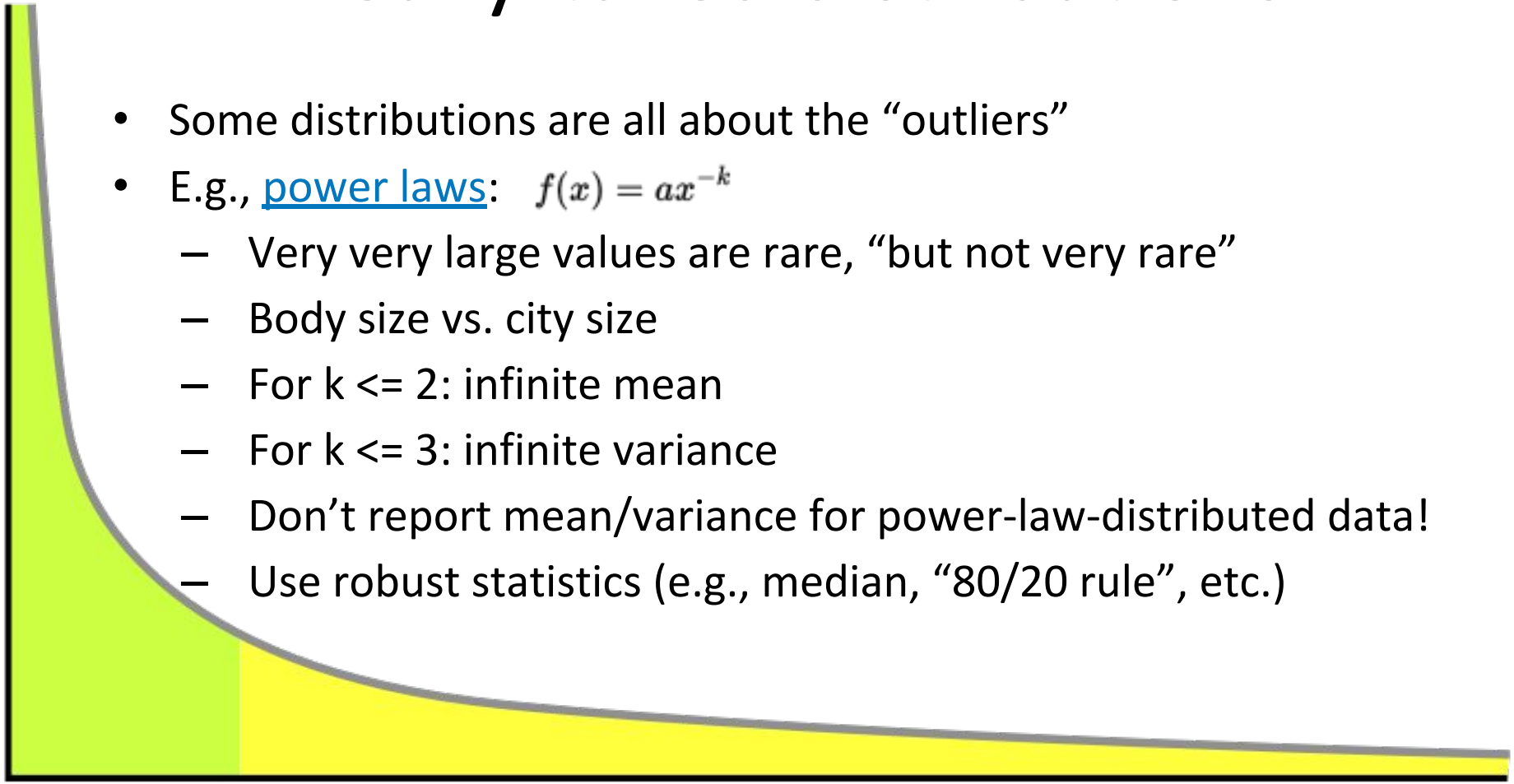
	year	stint	g	ab	r
count	100.00000	100.000000	100.000000	100.000000	100.00000
mean	2006.92000	1.130000	52.380000	136.540000	18.69000
std	0.27266	0.337998	48.031299	181.936853	27.77496
min	2006.00000	1.000000	1.000000	0.000000	0.00000
25%	2007.00000	1.000000	9.500000	2.000000	0.00000
50%	2007.00000	1.000000	33.000000	40.500000	2.00000
75%	2007.00000	1.000000	83.250000	243.750000	33.25000
max	2007.00000	2.000000	155.000000	586.000000	107.00000

Min, max, mean, std are not robust

Median, quartiles (and others) are robust

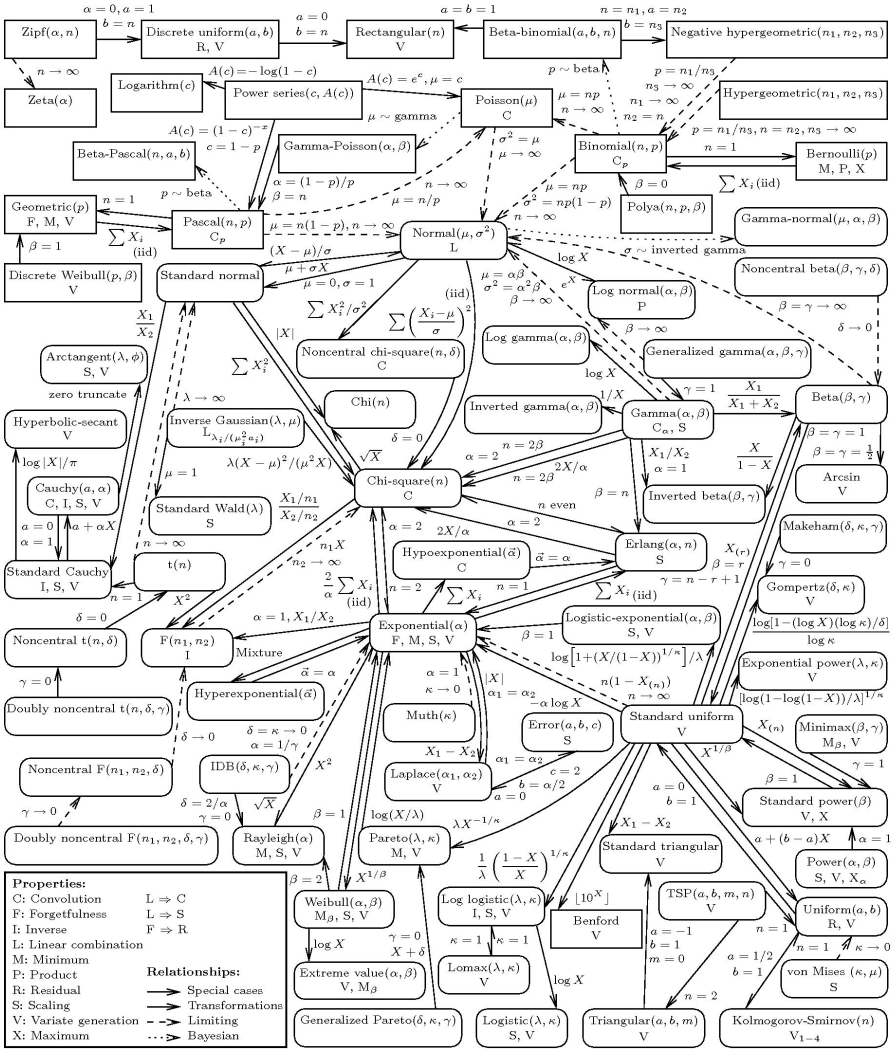
Check these Wikipedia pages

Heavy-tailed distributions



- Some distributions are all about the “outliers”
- E.g., power laws: $f(x) = ax^{-k}$
  - Very very large values are rare, “but not very rare”
  - Body size vs. city size
  - For k <= 2: infinite mean
  - For k <= 3: infinite variance
  - Don’t report mean/variance for power-law-distributed data!
  - Use robust statistics (e.g., median, “80/20 rule”, etc.)

Distributions



[link]

Distributions

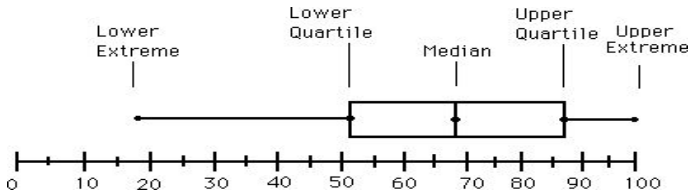
Some other important distributions:

- **Poisson:** the distribution of counts that occur at a certain “rate”; e.g., number of visits to a given website in a fixed time interval.
- **Exponential:** the interval between two such events.
- **Binomial/multinomial:** The number of counts of events (e.g., coin flips = heads) out of *n* trials.
- **Power-law/Zipf/Pareto/Yule:** e.g., frequencies of different terms in a document; city size

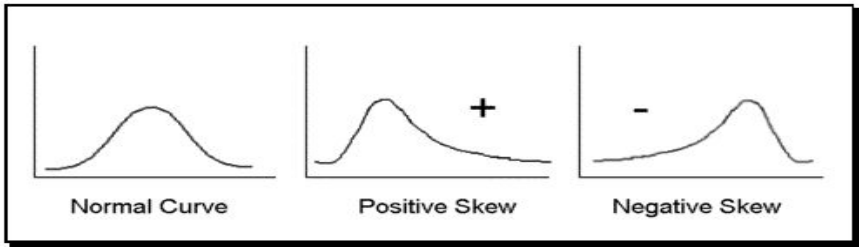
You should understand the distribution of your data before applying any model!

“Dear data, where are you from?”

- Visual inspection for ruling out certain distributions:
  e.g., when it’s asymmetric, the data cannot be normal. The histogram gives even more information.

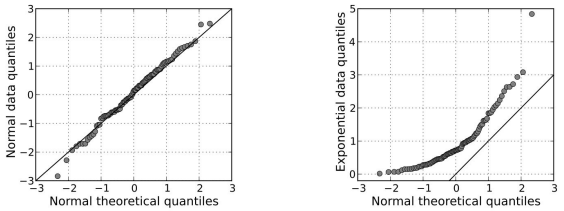


Box plot



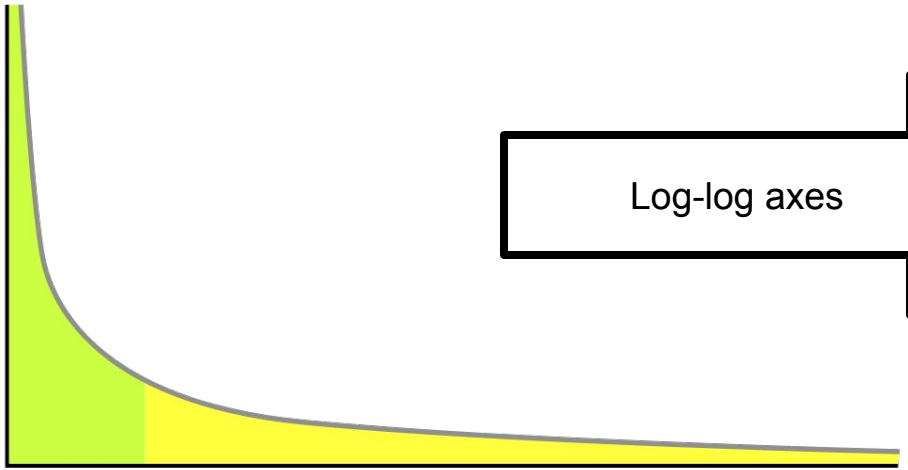
(Smoothed) histogram

- Statistical tests:
  - Goodness-of-fit tests
  - Kolmogorov-Smirnoff test
  - Normality tests

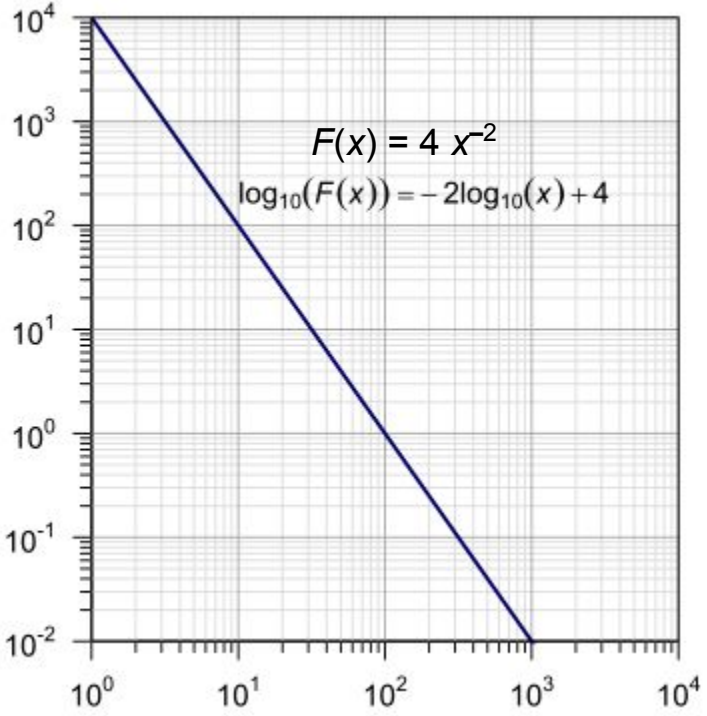


Quantile-quantile (QQ) plots

Recognizing a power law



Log-log axes



Sampling and uncertainty

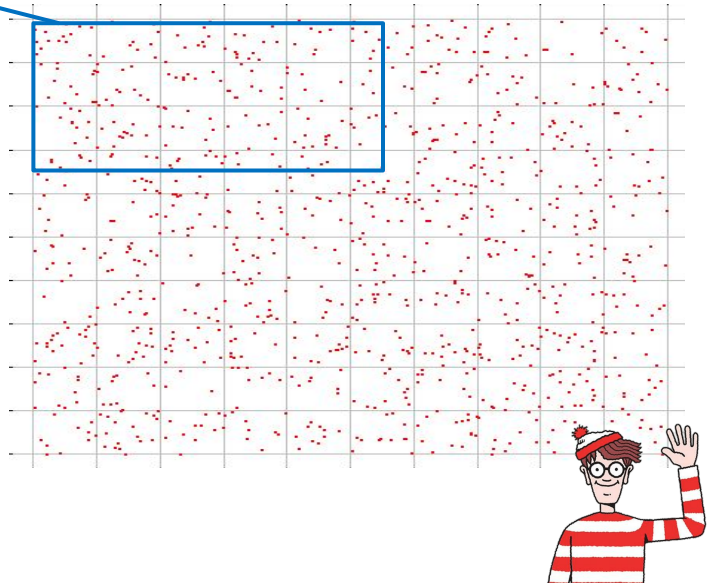
Measurement on samples

Datasets are **samples** from an **underlying distribution.**

We are most interested in **measures on the entire population,** but we have access only to a **sample** of it.

That makes measurement hard:

- Sample measurements have **variance**:
  variation between samples
- Sample measurements have **bias**:
  systematic variation from the population value.



Sampling is tricky

- Sometimes need to subsample for computational efficiency
  - Uniformly at random
  - Stratified sampling (e.g., equal number per quantile)
  - Importance sampling
- Sometimes need to subsample during data collection
  - E.g., Google Flu Trends: only Google users
  - Careful: bias!

Choice of Sample

In 1936 *Literary Digest* magazine sent more than 10 million ballots to readers to get their preferences in the upcoming presidential election between Franklin D. Roosevelt and Alfred M. Landon. The returns indicated that Landon, the Republican nominee, would win handily.

Roosevelt won in the first great landslide election of the 20th century. He carried all states except Maine and Vermont.

http://www.britannica.com/ebi/article-9276570?&query=newspapers&ct=%22ebi%22

Choice of Sample

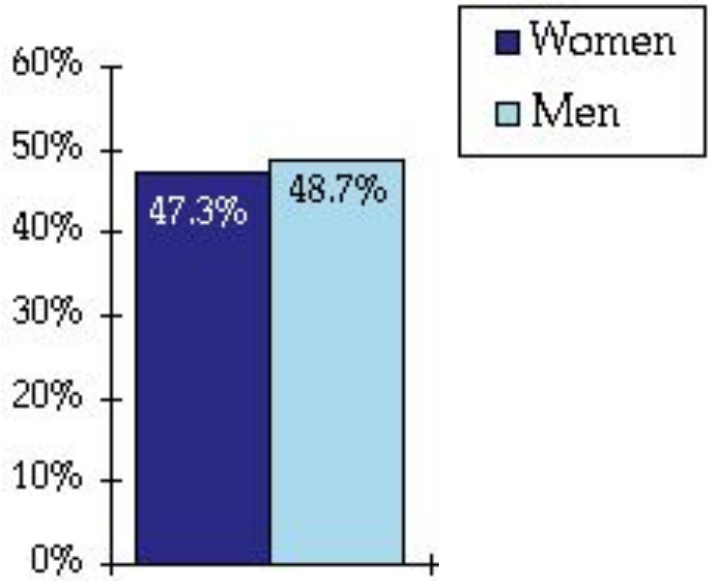
The Hawaii State Senate held hearings when it was considering a law requiring that motorcyclists wear helmets. Some motorcyclists testified that they had been in crashes in which helmets would not have been helpful. Which important group was unable to testify?

Taken from Triola, Essentials of Statistics. Addison Wesley.

So you have a biased dataset...

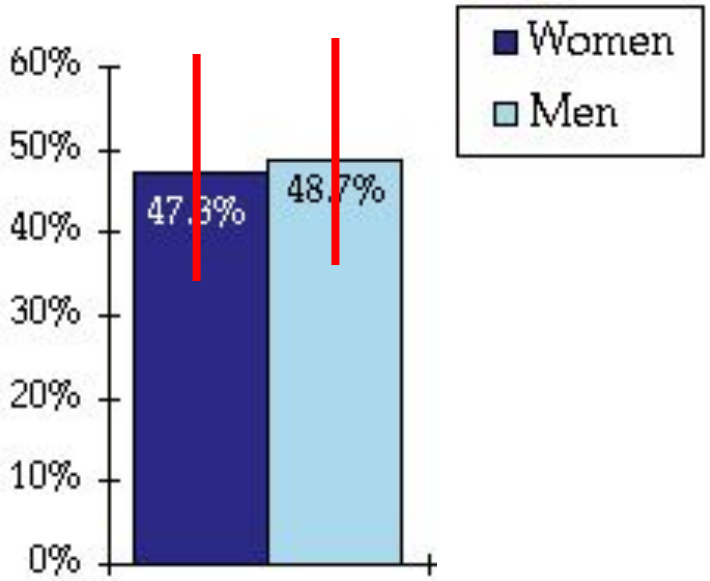
- “Found data”
- Observational studies
- Entire lecture on this (in 2 weeks from today)

Who likes Snickers better?



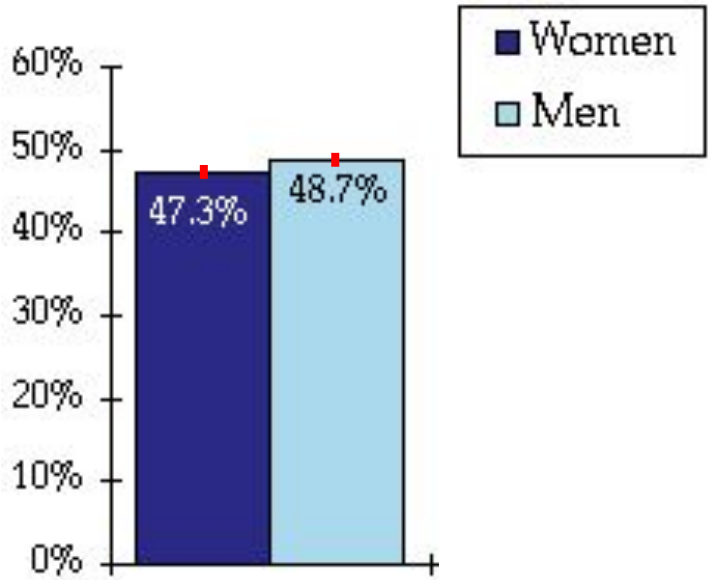
Who likes Snickers better?

Standard deviation in red



Who likes Snickers better?

Standard deviation in red

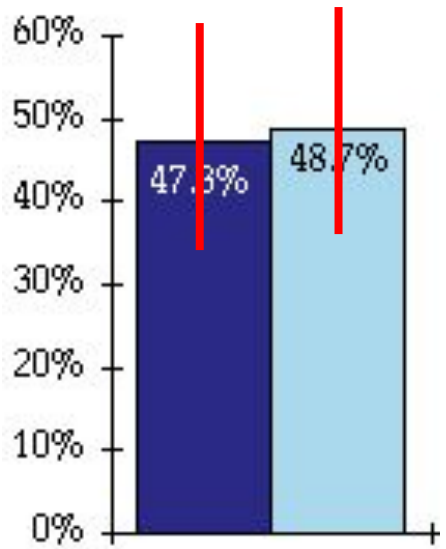


Quantifying uncertainty

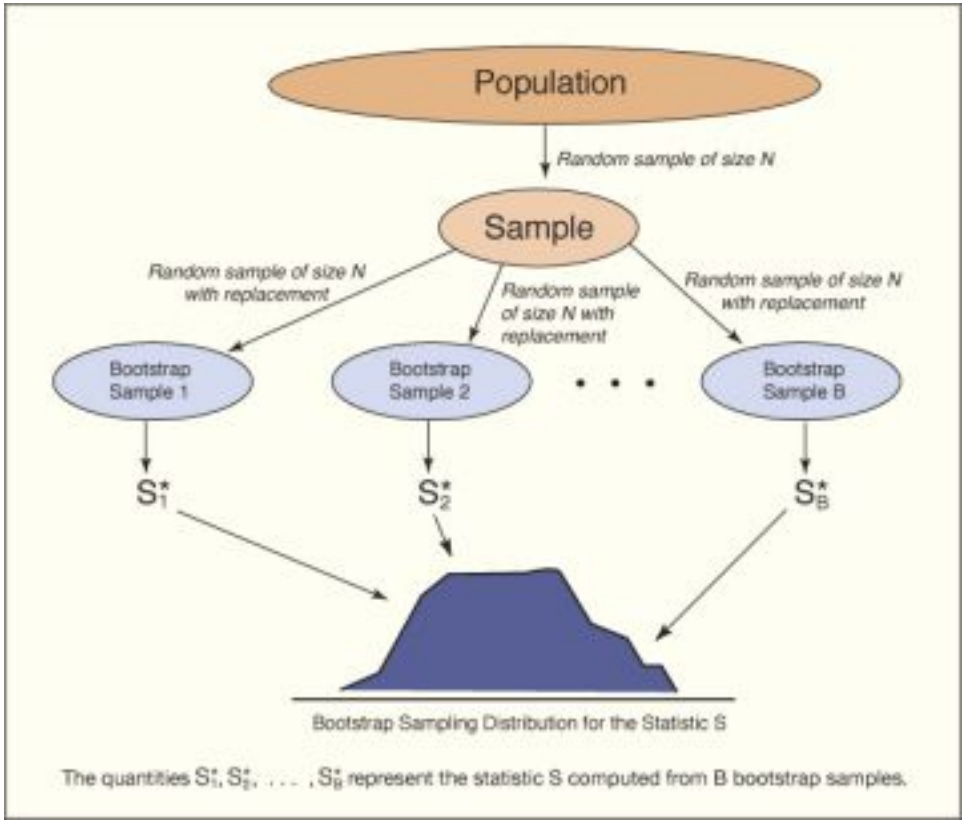
- Whenever you report a statistic, you need to quantify how certain you are in it!
- Even a complete dataset is a sample
- Entire field: hypothesis testing (later today)
- All plots should have error bars!

Error bars

- Can represent many things
  (→ always explain; always question):
  - Standard deviation
  - Standard error of the mean:
    sd/sqrt(n)
  - Confidence intervals
    - Parametric (ugh!)
    - Non-parametric (yay!): bootstrap resampling



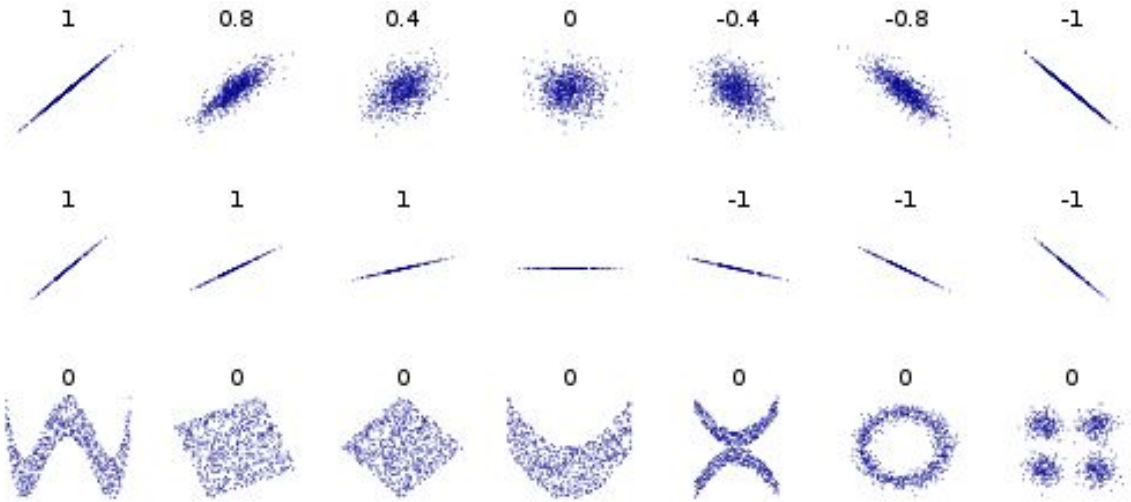
Bootstrap resampling



Relating two variables

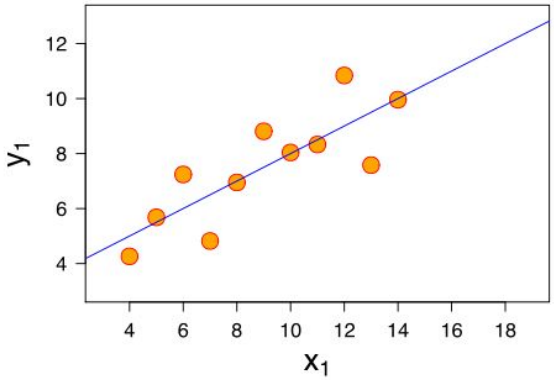
Pearson's correlation coefficient

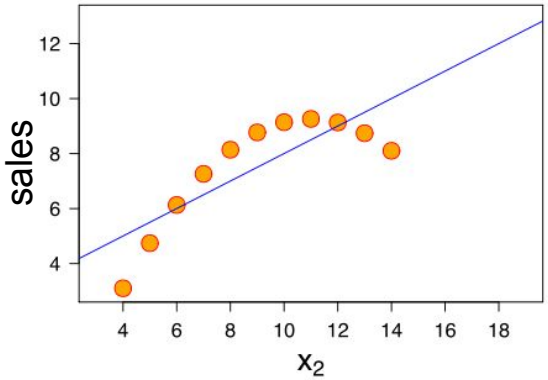
- "Amount of linear dependence"

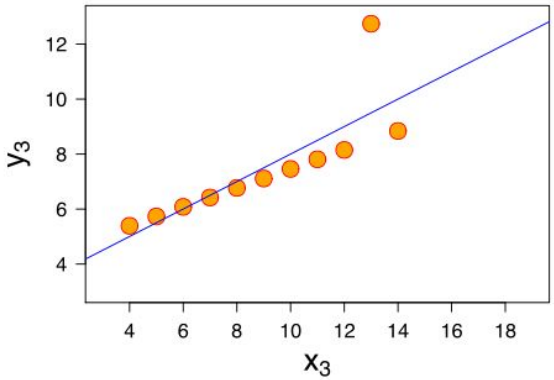


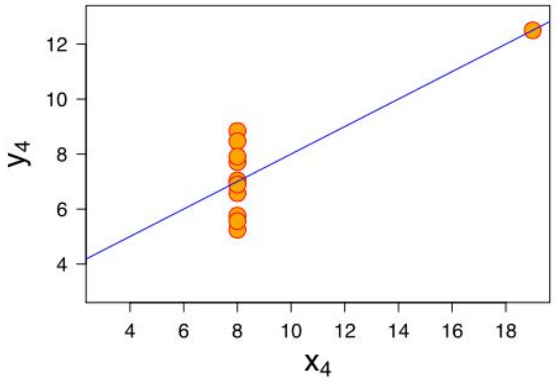
- More general:
  - Rank correlation, e.g., Spearman's correlation coefficient
  - Mutual information

Anscombe's quartet

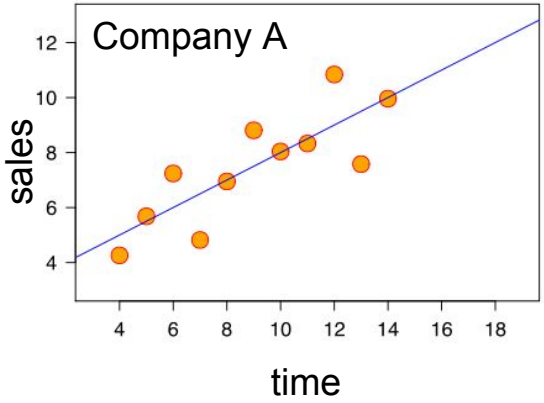


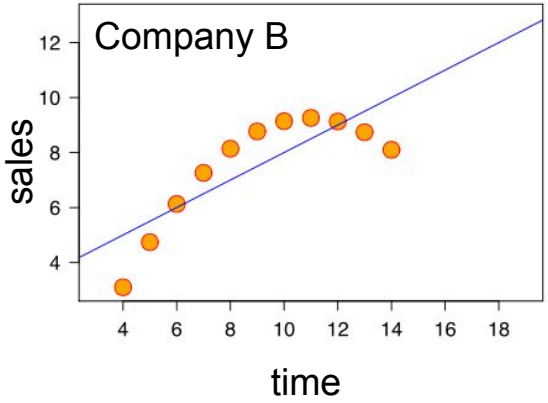


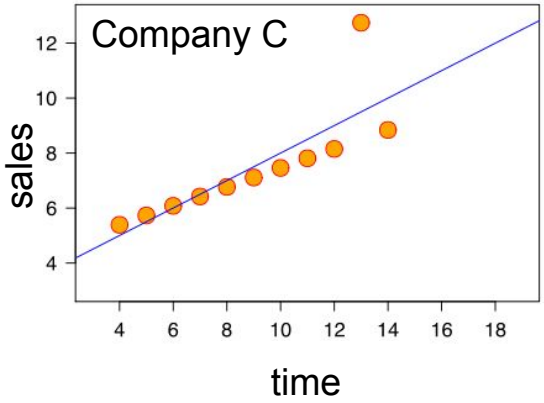




Anscombe's triplet

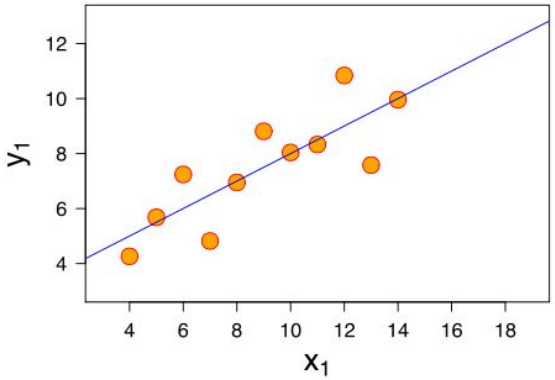


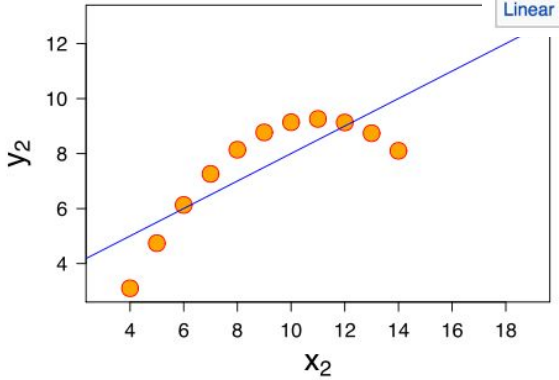


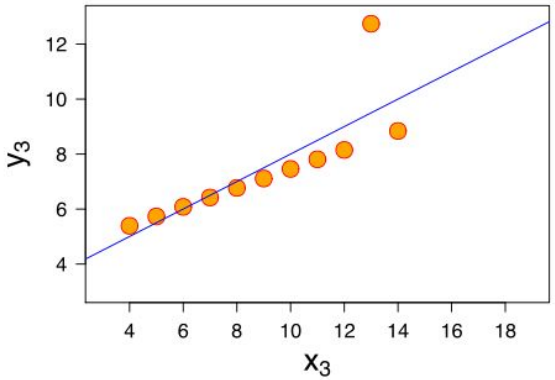


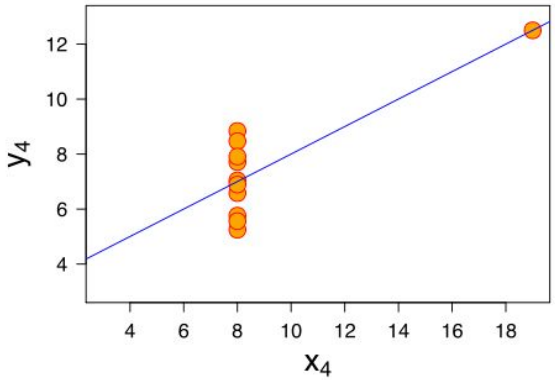
Anscombe's quartet

Property	Value
Mean of x in each case	9 (exact)
Sample variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Sample variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)









Property	Value
Mean of x in each case	9 (exact)
Sample variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Sample variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

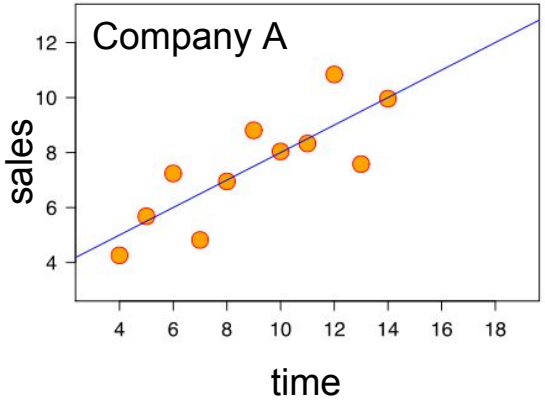
Anscombe's quartet

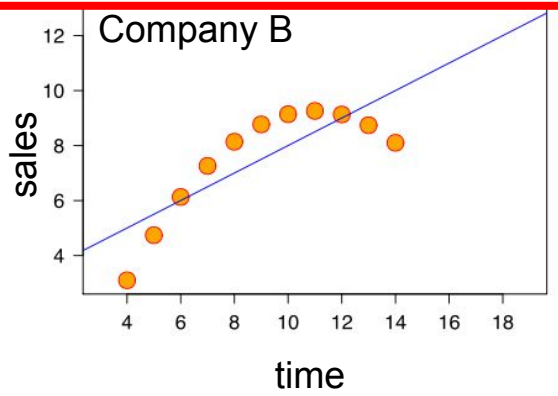
Illustrates the **importance of looking at a set of data graphically** before starting to analyze

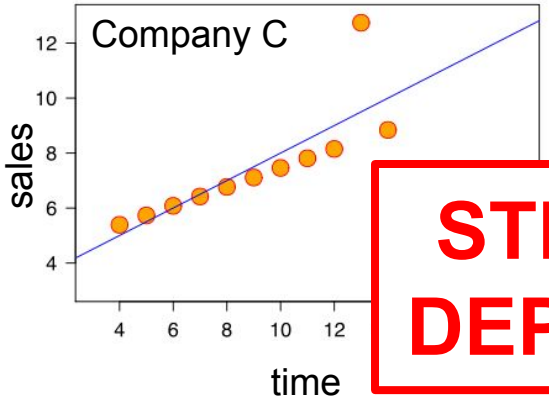
Highlights the *inadequacy of basic statistical properties for describing realistic datasets*

More on Wikipedia

WEAKEST DEPENDENCE







STRONGEST DEPENDENCE

Good solutions:

- Visual reasoning
- Spearman rank correlation (rather than Pearson correlation)

Incorrect solutions:

- "No difference"
- Pearson correlation
- Slope of linear regression

(but nearly identical for all plots)

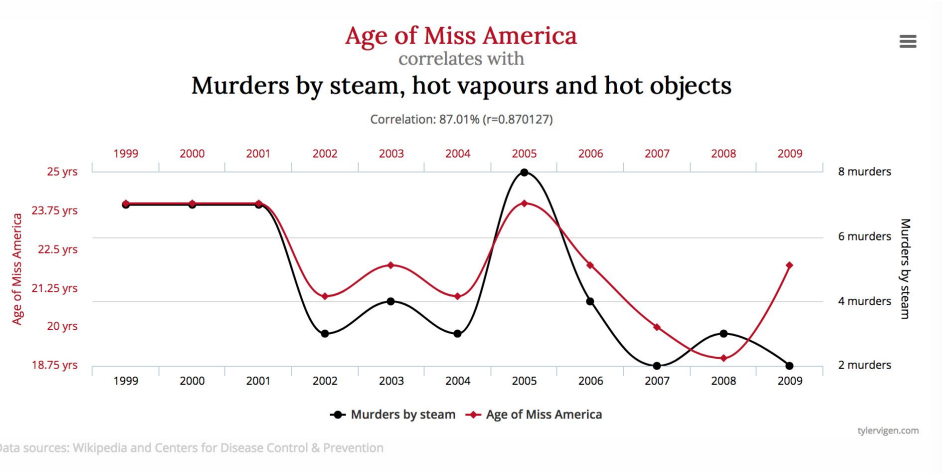
- median(sales)
- stdev(sales)
- max(sales) - min(sales)
- avg(sales/time)

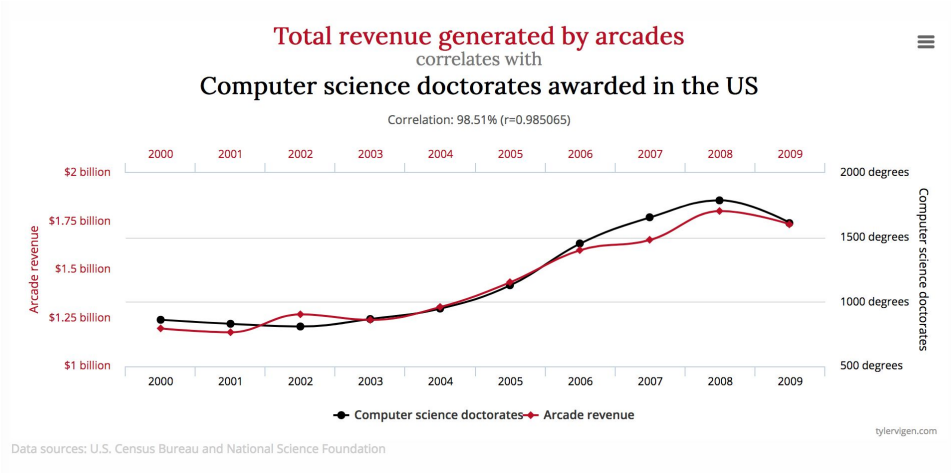
(not taking order into account)

- avg((sales - time)^2)
- (avg(sales) - avg(time))^2
- "Outliers imply lower dependence"
- "Quadratic dependence (B) counts more than linear dependence"

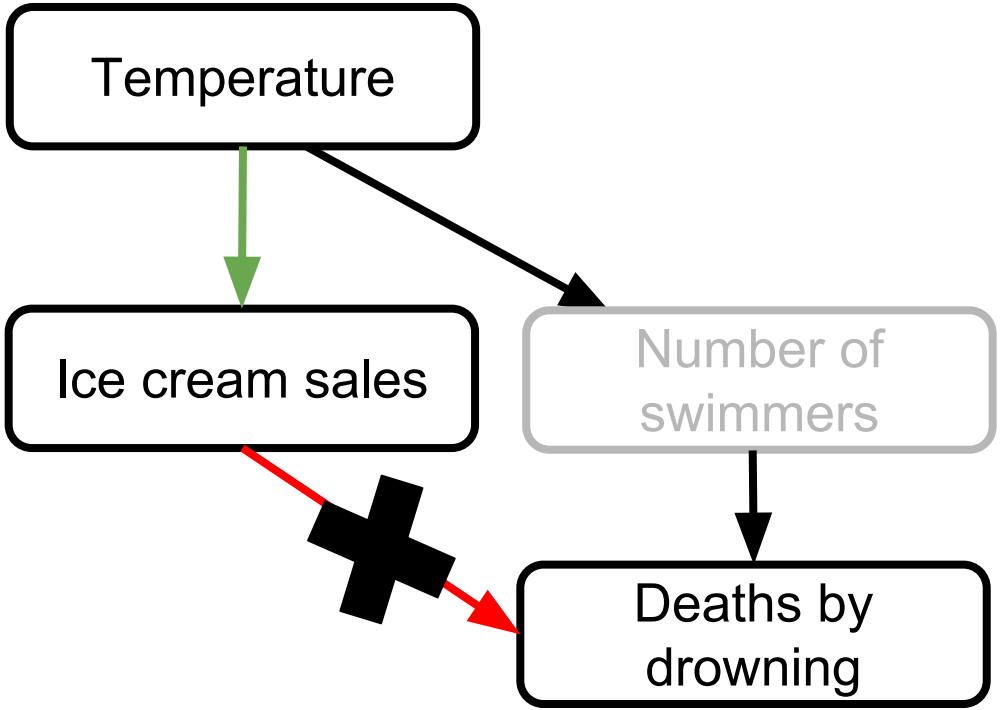
Correlation coefficients are tricky!

- http://guessthecorrelation.com/
- Correlation != causation (cf. lecture in 2 weeks)
  http://www.tylervigen.com/spurious-correlations





Ice cream sales vs. deaths by drowning, anyone?



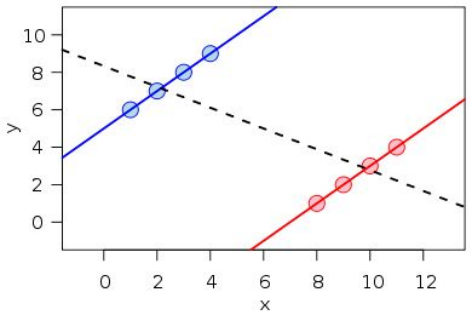
UC Berkeley gender bias (**?**)

Admission figures from 1973

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

Simpson's paradox



When a trend appears in different groups of data **but disappears or reverses when these groups are combined** -- beware of aggregates!

From the previous example, **women tended to apply to competitive departments** with low rates of admission

Hypothesis testing

Rhine's paradox

Joseph Rhine was a parapsychologist in the 1950's (founder of the *Journal of Parapsychology* and the *Parapsychological Society, an affiliate of the AAAS*).

He ran an experiment where subjects had to guess whether 10 hidden cards were red or blue.

He found that about 1 person in 1000 had ESP (Extrasensory perception), i.e. they could guess the color of all 10 cards!

Q: Do you agree?

Rhine's paradox

He called back the "psychic" subjects and had them do the same test again. They all failed.

He concluded that **the act of telling psychics that they have psychic abilities** causes them to lose them…

Which of the following statements about p-values are true?

1	*If P = .05, the null hypothesis has only a 5% chance of being true.*
2	*A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.*
3	*A statistically significant finding is clinically important.*
4	*Studies with P values on opposite sides of .05 are conflicting.*
5	*Studies with the same P value provide the same evidence against the null hypothesis.*
6	*P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.*
7	*P = .05 and P ≤.05 mean the same thing.*
8	*P values are properly written as inequalities (eg, "P ≤.02" when P = .015)*
9	*P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.*
10	*With a P = .05 threshold for significance, the chance of a type I error will be 5%.*
11	*You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.*
12	*A scientific conclusion or treatment policy should be based on whether or not the P value is significant.*

Preliminaries

- A huge subject; can take entire classes on it
- A black art; many people hate it
- cf. Bayesian vs. frequentist debate (a.k.a. war)
- Need to understand basics even if you don't use it yourself
- Never use it without understanding exactly what you're doing

Hypothesis testing

Reasoning:

- Flip a coin 100 times; 40 heads; “Is the coin fair?”
- Null hypothesis: “yes”; alternative hypothesis: “no”
- “How likely would I be to see 40 or fewer heads if the null hypothesis were true?”
- If this probability is large, the null hypothesis suffices (and is thus not rejected)
- Otherwise, keep experimenting

Hypothesis testing

- Idea: Gain (weak and indirect) support for a hypothesis H_A by **ruling out a null hypothesis H_0**

- A **test statistic** is some measurement we can make on the data which is likely to be **large under H_A** but **small under H_0**

Hypothesis testing

- Gain (weak and indirect) support for a hypothesis H_A **(the coin is biased)** by means of disproving a null hypothesis H_0 **(the coin is fair)**.
- A test statistic is some measurement we can make on the data which is likely to be large under H_A but small under H_0.
  the number of heads after *k* coin tosses (one sided)
  the difference between number of heads and *k*/2 (two-sided)

- **Note:** tests can be either one-tailed or two-tailed. Here a two-tailed test is convenient because it treats very large and very small counts of heads the same way.

Hypothesis testing

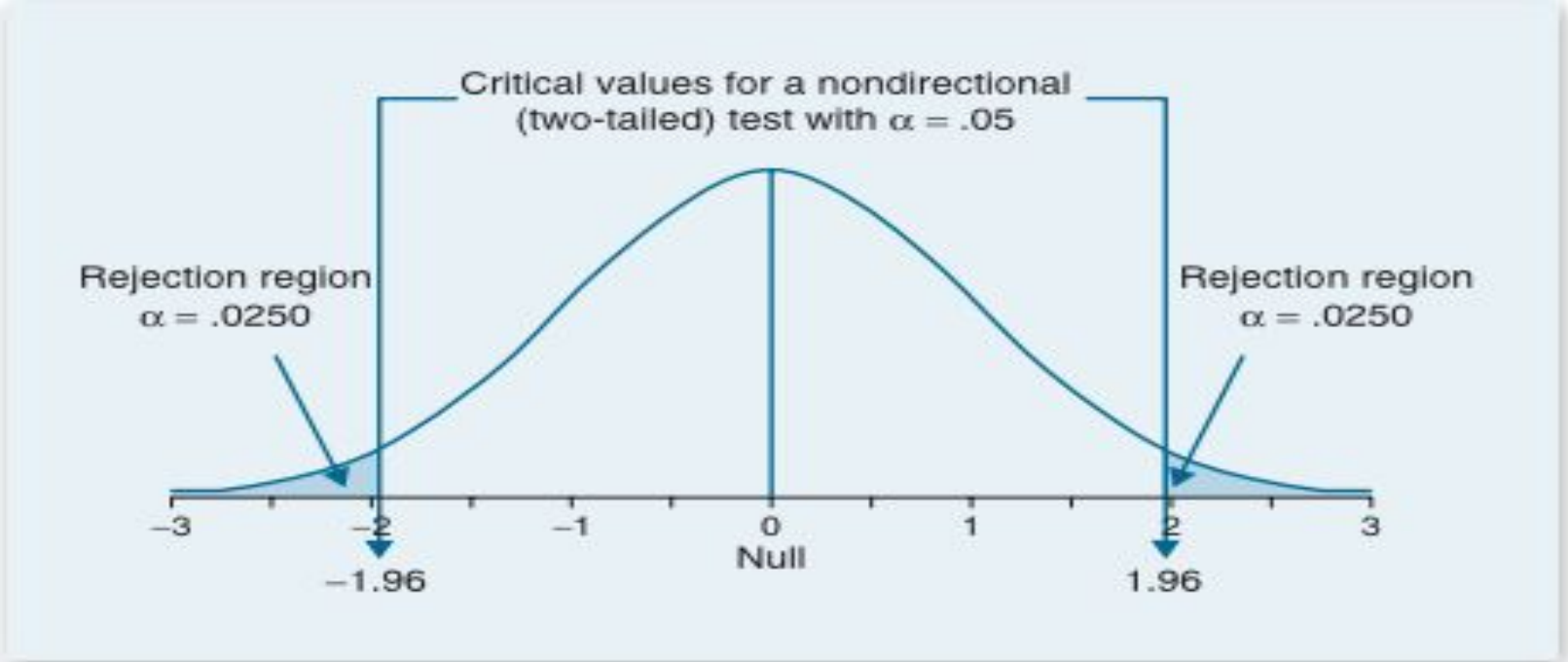
- Another example:
  - Two samples *a* and *b*, normally distributed, from *A* and *B*.
  - Null hypothesis H_0: mean(*A*) = mean(*B*)
    test statistic is: *s* = mean(*a*) – mean(*b*).
  - Under H_0, *s* has mean zero and is normally distributed*.
  - But it is “large” if the two means are different.

* Because the sum of two independent, normally distributed variables is also normally distributed.

Hypothesis testing

- s = mean(a) – mean(b) is our test statistic;
  the null hypothesis H_0 is "mean(A) = mean(B)"
  - We reject H_0 if $Pr(S > s \mid H_0) < \alpha$, i.e., if the probability of a statistic value **at least as large as** *s* is small.
  - $Pr(S > s \mid H_0)$ is the infamous **p-value;** α is called **significance level**
  - α is a suitable "small" probability, say 0.05.
  - α directly controls the false-positive rate (probability of rejecting H_0 although it is true): higher $\alpha \rightarrow$ higher false-positive rate
  - As we make α smaller, the false-negative rate increases (probability of not rejecting H_0 although it is false)
  - Common values for α: 0.05, 0.02, 0.01, 0.005, 0.001

Two-tailed significance



When the p-value is less than 5% (p < .05), we reject the null hypothesis

Select your tests

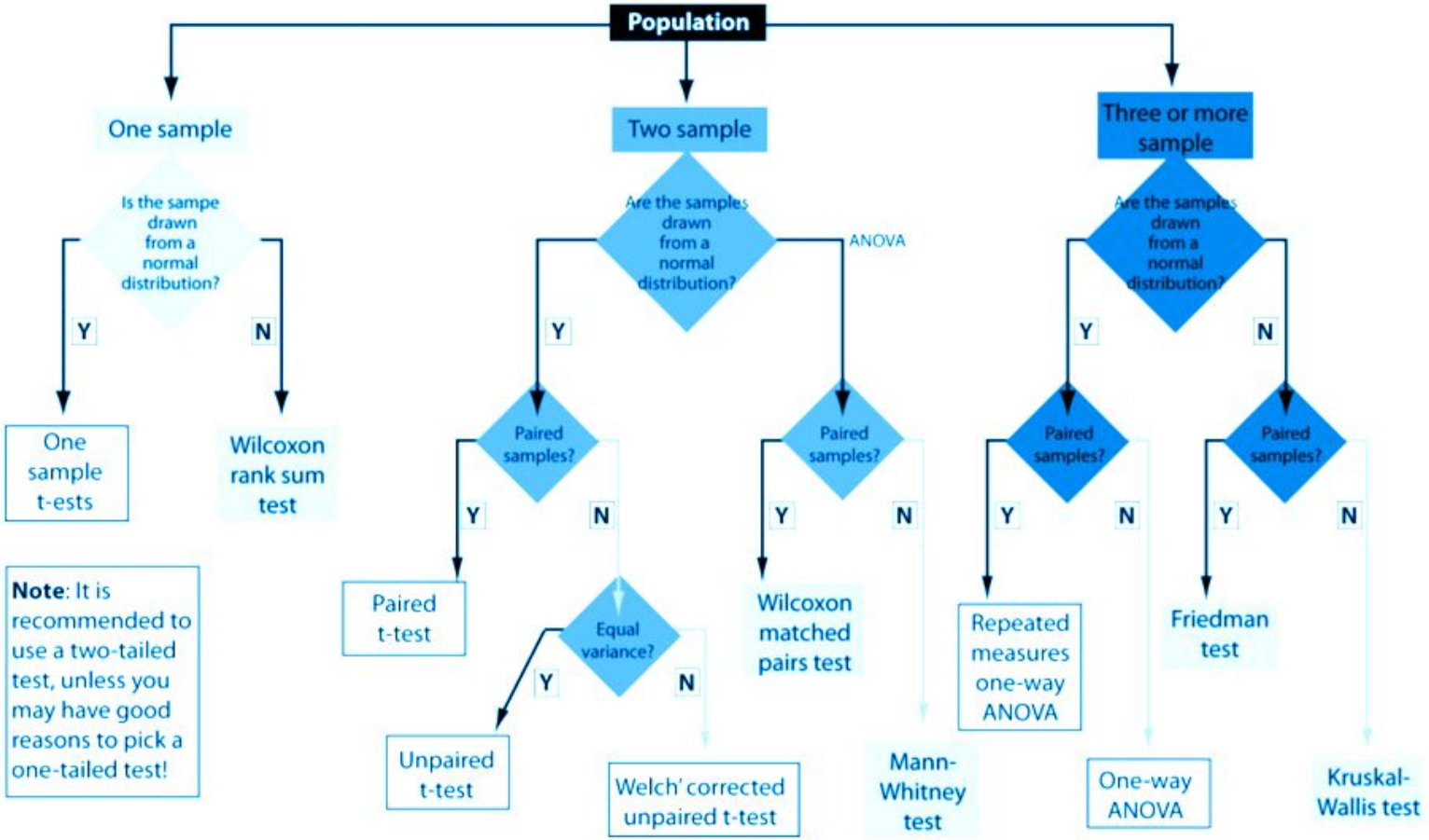
Testing is a bit like **finding the right recipe** based on these ingredients:

i) Question, ii) Data type, iii) Sample size, iv) Variance known?, v) Variance of several groups equal?

Good news: **Plenty of tables available!**

http://www.ats.ucla.edu/stat/mult_pkg/whatstat/default.htm (with examples in R, SAS, Stata, SPSS)

http://sites.stat.psu.edu/~ajw13/stat500_su_res/notes/lesson14/images/summary_table.pdf


Population
One sample
Two sample
Three or more sample
Is the sampe drawn from a normal distribution?
Are the samples drawn from a normal distribution?
Are the samples drawn from a normal distribution?
ANOVA
Y
N
One sample t-ests
Wilcoxon rank sum test
Paired samples?
Paired t-test
Equal variance?
Unpaired t-test
Welch' corrected unpaired t-test
Paired samples?
Wilcoxon matched pairs test
Mann-Whitney test
Paired samples?
Repeated measures one-way ANOVA
One-way ANOVA
Paired samples?
Friedman test
Kruskal-Wallis test
Note: It is recommended to use a two-tailed test, unless you may have good reasons to pick a one-tailed test!

p-values

- Widely used in all sciences
- They are widely misunderstood!
- Don’t dare to use them if you don’t understand them! (example)
- Large p means that even under a null hypothesis your data would be quite likely
- This tells you nothing about the alt. hypothesis

p-values

- Historically, not meant as a method for formally deciding whether a hypothesis is true or not
- Rather, an informal tool for assessing a particular result
- Low p-value means: the simple null hypothesis doesn't explain the data, so keep looking for other explanations!
- p = 0.05 means: if you repeat experiment 20 times, you'll see extreme data even under null hypothesis → you might have "lucked out"
- Look at the y-axis, not just the p-value!

p-values

- Important to understand what p-values are
- Maybe even more important to understand what they are not…
- Read this paper: A Dirty Dozen: 12 P-Value Misconceptions

Table 1 Twelve *P*-Value Misconceptions

1	*If P = .05, the null hypothesis has only a 5% chance of being true.*
2	*A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.*
3	*A statistically significant finding is clinically important.*
4	*Studies with P values on opposite sides of .05 are conflicting.*
5	*Studies with the same P value provide the same evidence against the null hypothesis.*
6	*P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.*
7	*P = .05 and P ≤.05 mean the same thing.*
8	*P values are properly written as inequalities (eg, "P ≤.02" when P = .015)*
9	*P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.*
10	*With a P = .05 threshold for significance, the chance of a type I error will be 5%.*
11	*You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.*
12	*A scientific conclusion or treatment policy should be based on whether or not the P value is significant.*

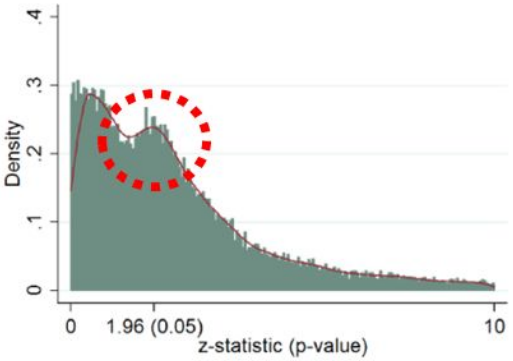
“p-value hacking”

1. Stop collecting data once $p<.05$
2. Analyze many measures, but report only those with $p<.05$.
3. Collect and analyze many conditions, but only report those with $p<.05$.
4. Use covariates to get $p<.05$.
5. Exclude participants to get $p<.05$.
6. Transform the data to get $p<.05$.

Economics

Brodeur et al (AEJ:A, in press)

“Star Wars: The empirics strike back”

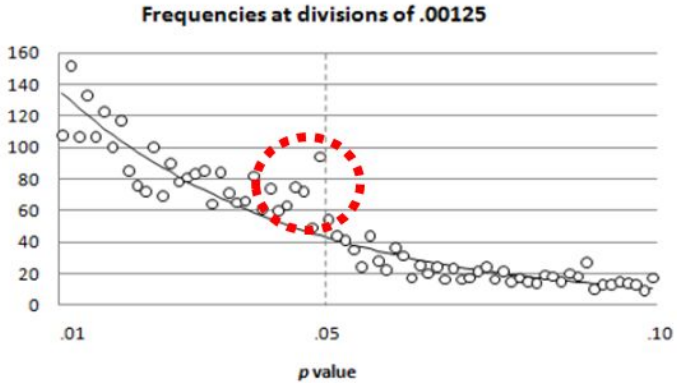


(b) De-rounded distribution of z-statistics.

Psychology

Masicampo Lalande (QJEP, 2012)

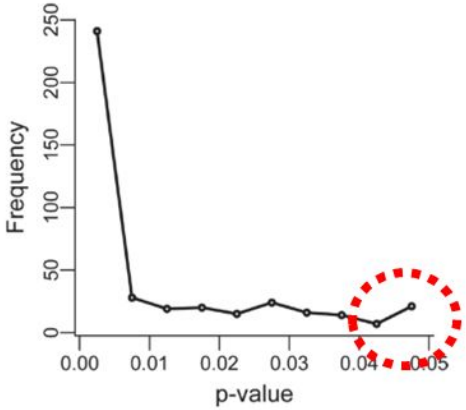
“A peculiar prevalence of p values just below .05”



Biology

Head et al (PLOS Biology 2015)

“Extent and Consequences of P-Hacking in Science”



BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Routledge
Taylor & Francis Group

Editorial

David Trafimow and Michael Marks
New Mexico State University

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

With the banning of the NHSTP from BASP, what are the implications for authors? The following are

a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that when in a state of ignorance, the researcher should assign an equal probability to each possibility. The

Alternative approach: Bayes factors

$$\frac{\text{Prob(Data, under } H_0)}{\text{Prob(Data, under } H_A)}$$

- See here
- Great (and amusing) explanation of difference between hypothesis-testing approach and Bayesian approach: Chapter 37 in MacKay's book on "Information Theory, Inference, and Learning Algorithms"

Multiple-hypothesis testing

If you perform experiments over and over, you're bound to find something

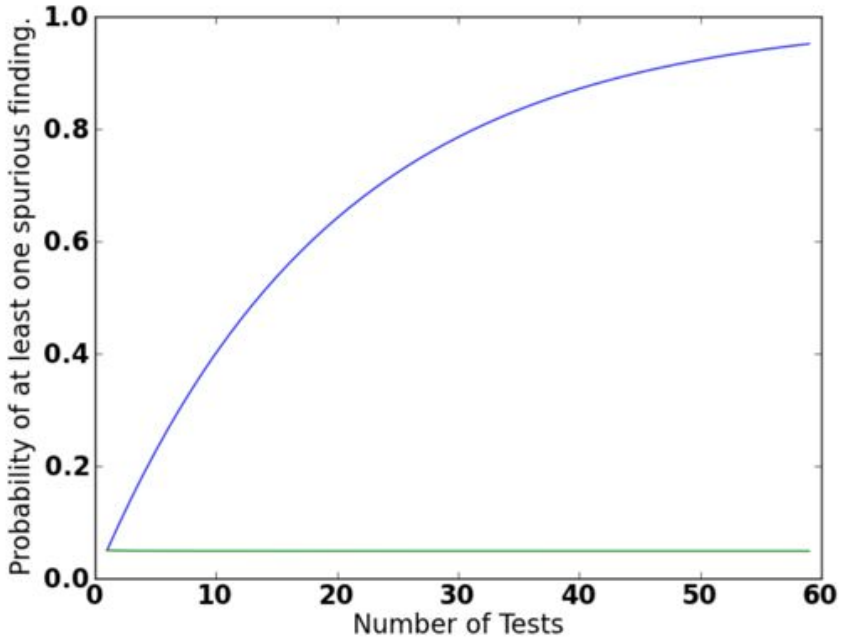
Significance level must be adjusted down when performing multiple hypothesis tests!

P(detecting an effect when there is none) = $\alpha = 0.05$

P(detecting no effect when there is none) $= 1 - \alpha$

P(detecting no effect when there is none, on every experiment) $= (1 - \alpha)^k$

P(detecting an effect when there is none on at least one experiment) $= 1 - (1 - \alpha)^k$



$\alpha = 0.05$

"Familywise Error Rate"

Family-wise Error Rate Corrections

Bonferroni Correction

- Just divide by the number of hypotheses

$$\alpha_c = \frac{\alpha}{k}$$

Šidák Correction

- Asserts independence

$$\alpha = 1 - (1 - \alpha_c)^k$$

$$\alpha_c = 1 - (1 - \alpha)^{\frac{1}{k}}$$

There are three kinds of lies: lies, damned lies, and statistics.

Benjamin Disraeli (1804 - 1881)

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2018-lec5-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- Where is Waldo?
- …

Non-Parametric Tests

All the tests so far are parametric tests that assume the data are **normally distributed**, and that the samples are **independent of each other and all have the same distribution** (IID).

They may be arbitrarily inaccurate if those assumptions are not met.

Always make sure your data satisfies the assumptions of the test you're using, i.e. watch out for:

- **Outliers** – will corrupt many tests that use variance estimates.
- **Correlated values as samples**, e.g. if you repeated measurements on the same subject.
- **Skewed distributions** – give invalid results.

Non-parametric tests

These tests make no assumption about the distribution of the input data, and can be used on very general datasets:

- **K-S test** (today)
- *Permutation tests* and *Bootstrap confidence intervals* (we will see them in the following lectures)

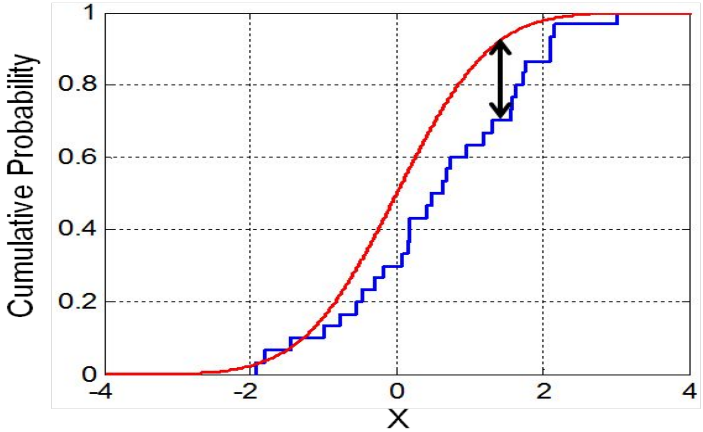
K-S test

The K-S (Kolmogorov-Smirnov) test is a very useful test for checking whether two (continuous or discrete) distributions are the same.

In the **one-sided test**, an observed distribution (e.g. some observed values or a histogram) is compared against a reference distribution (e.g., power-law)

In the **two-sided test**, two observed distributions are compared.

The K-S statistic is just the **max distance between the CDFs** of the two distributions.



K-S test

The K-S test can be used to test **whether a data sample has a normal distribution** or not.

Thus it can be used as a sanity check for any common parametric test (which assumes normally-distributed data).

It can also be used to compare distributions of data values in a large data pipeline: **Most errors will distort the distribution of a data parameter and a K-S test can detect this**.