Introduction to Data Science and Analytics (DSC510)

Describing Data

George Pallis

Overview of today's lecture

- Part 1: Descriptive statistics
- Part 2: Quantifying uncertainty
- Part 3: Relating two variables

This course won't cover the basics of stats!

You know these things from prerequisite courses

But stats are a key ingredient of data analysis

Today: some highlights and common pitfalls

Part 1
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:

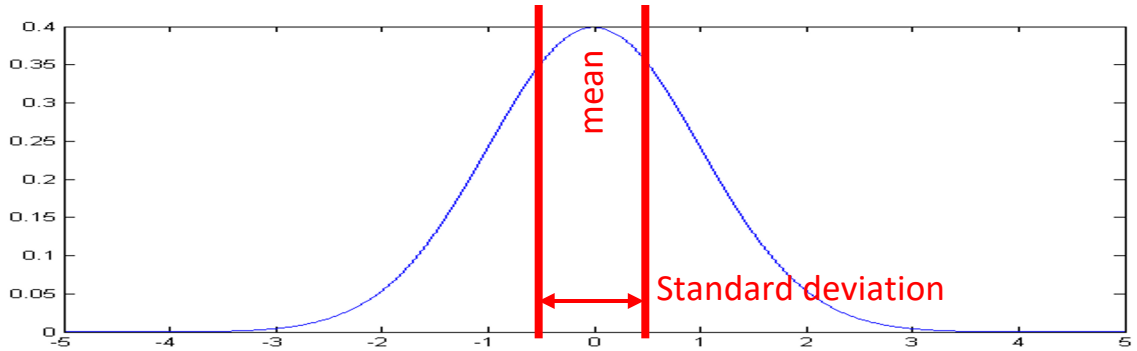
$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

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

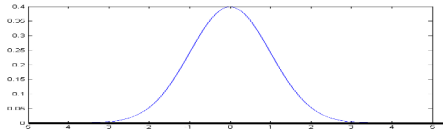
The normal distribution is completely characterized by mean and std.

```
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



x

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

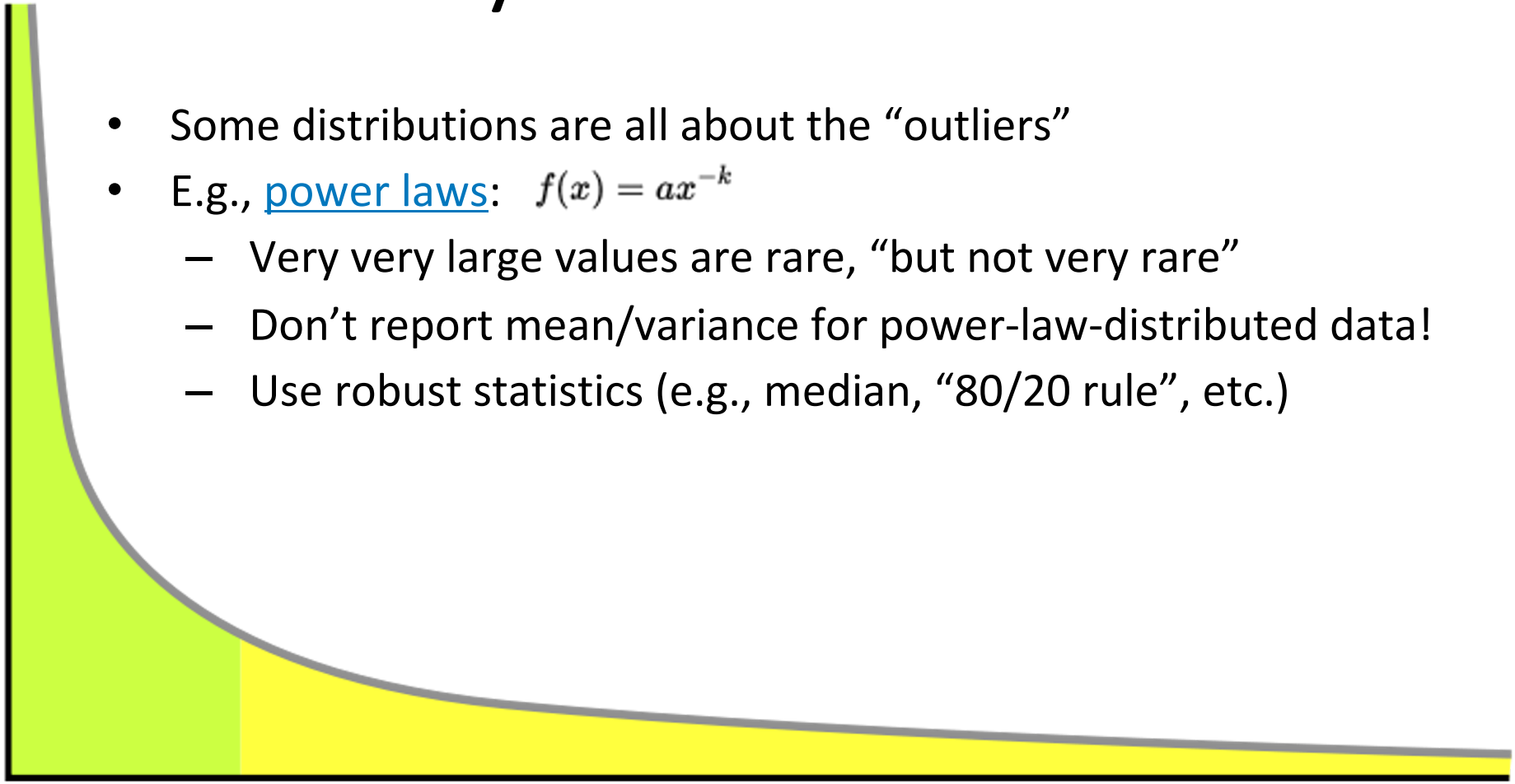
Min, max, mean, std are not robust

Median, quartiles (and others) are robust

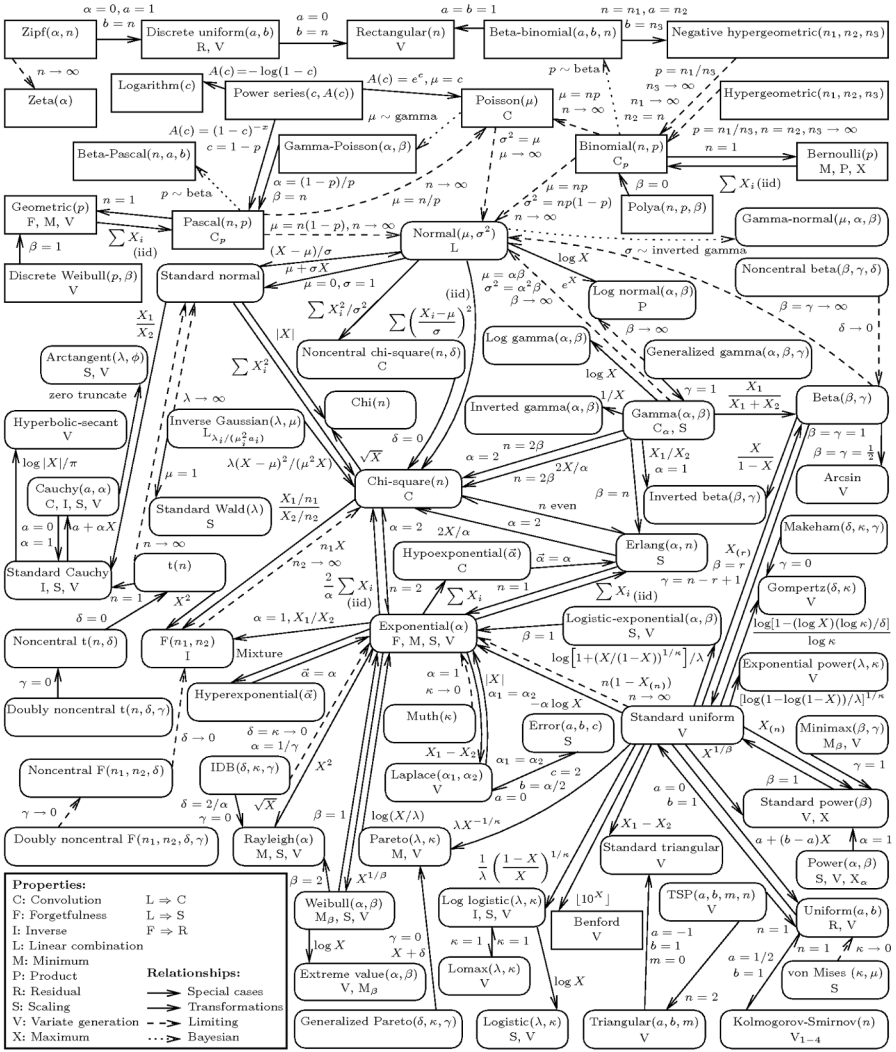
Check these Wikipedia pages

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”
  - Don’t report mean/variance for power-law-distributed data!
  - Use robust statistics (e.g., median, “80/20 rule”, etc.)



Distributions



[link]

Distributions

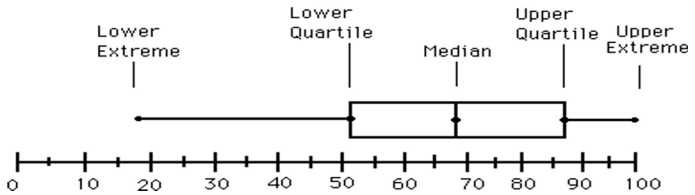
Some important distributions:

- **Normal:** a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. **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

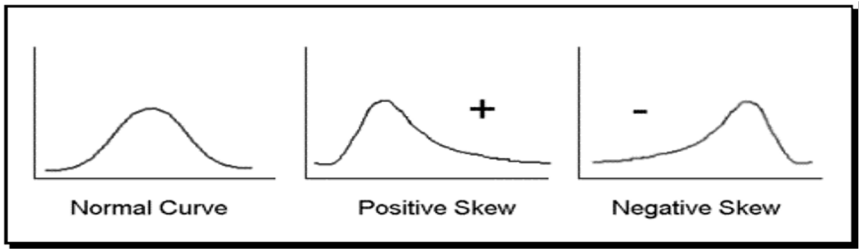
You should understand the distribution of your data before applying any model!

“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.

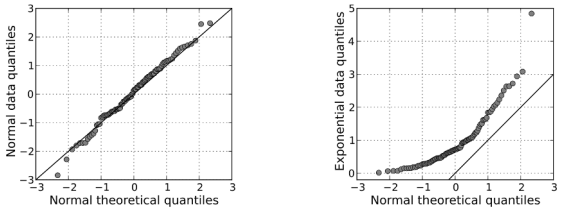


Box plot



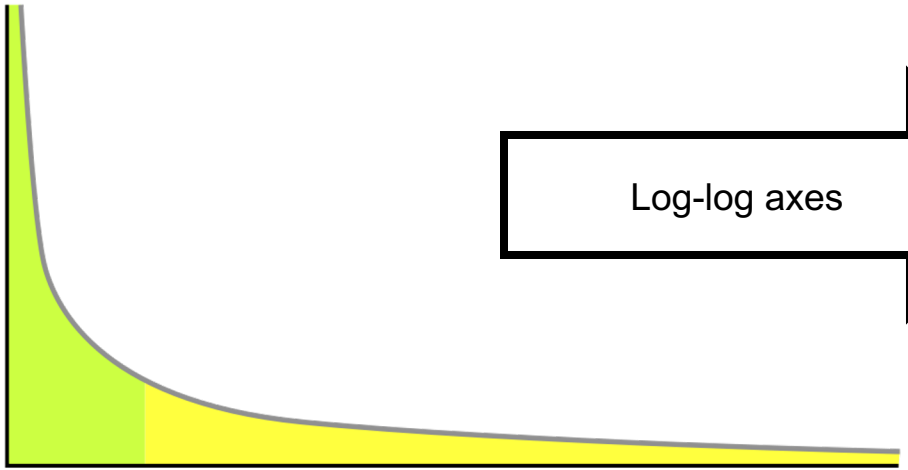
(Smoothed) histogram

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

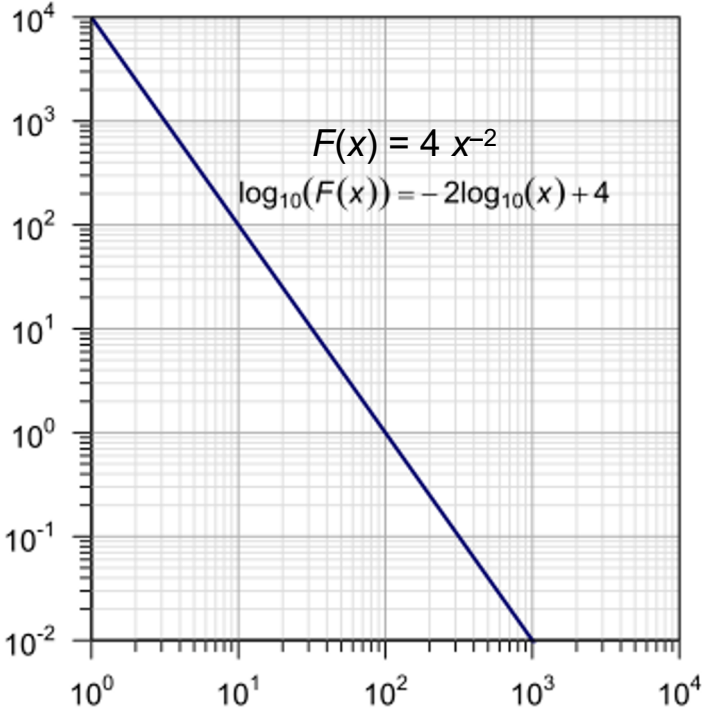


Quantile-quantile (QQ) plots

Recognizing a power law



Log-log axes



Part 2
Quantifying 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.

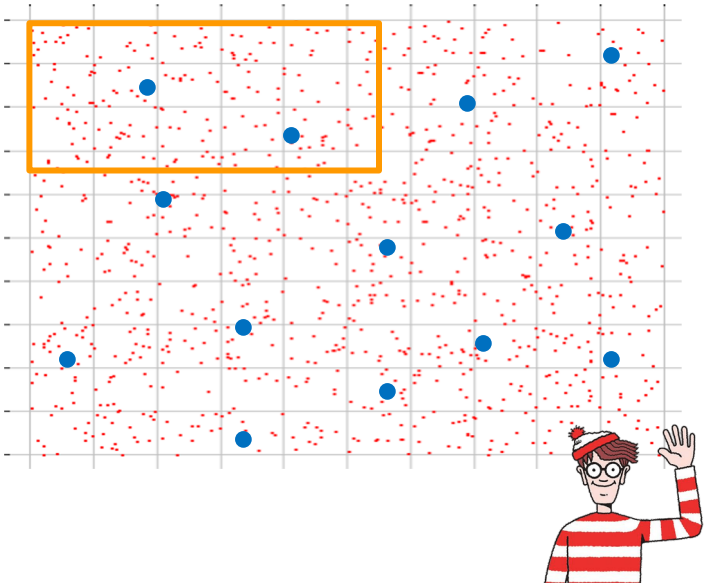
That makes measurement hard:

- Sample measurements have **variance**:

  variation between samples

- Sample measurements have **bias**:

  systematic variation from the population value.

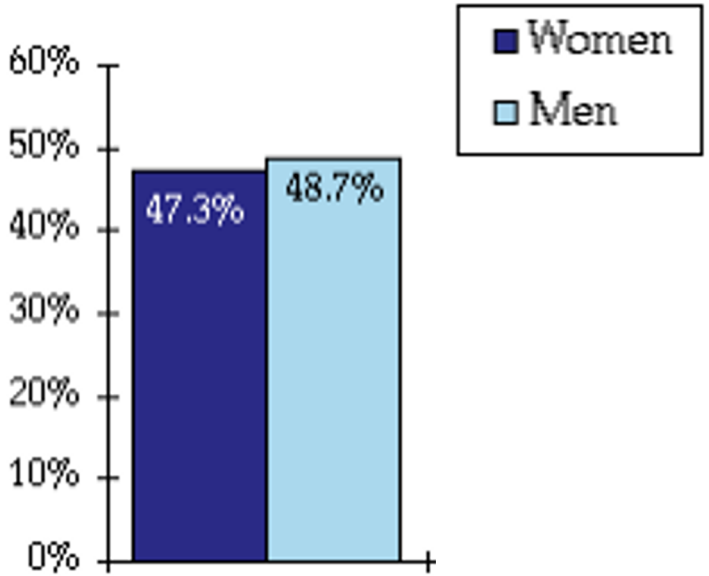


So you have a biased dataset…

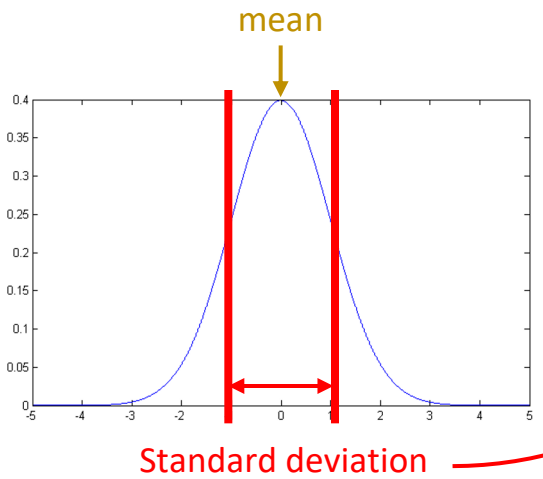
- “Found data”
- Observational studies
  - Entire lecture on this

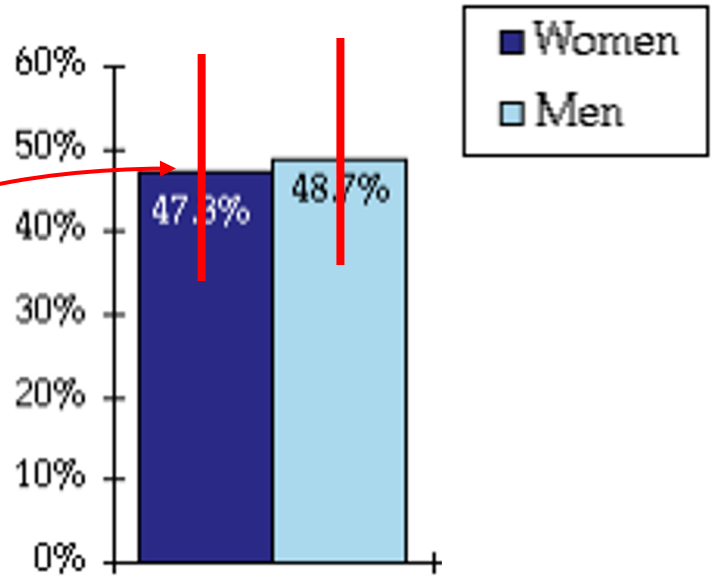
Who likes Snickers better?

- Most straightforward descriptive statistic to answer this question:
- Mean for each group (women, men)

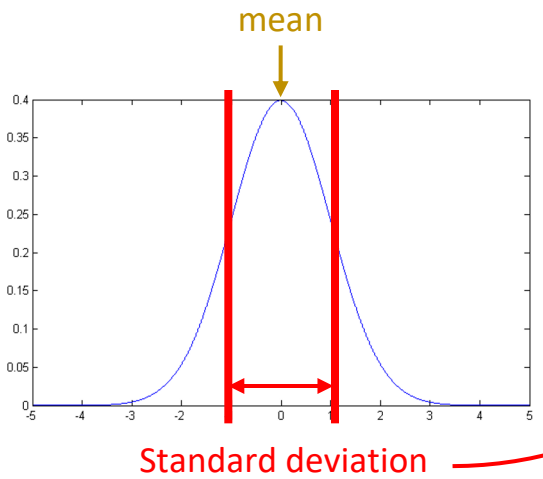


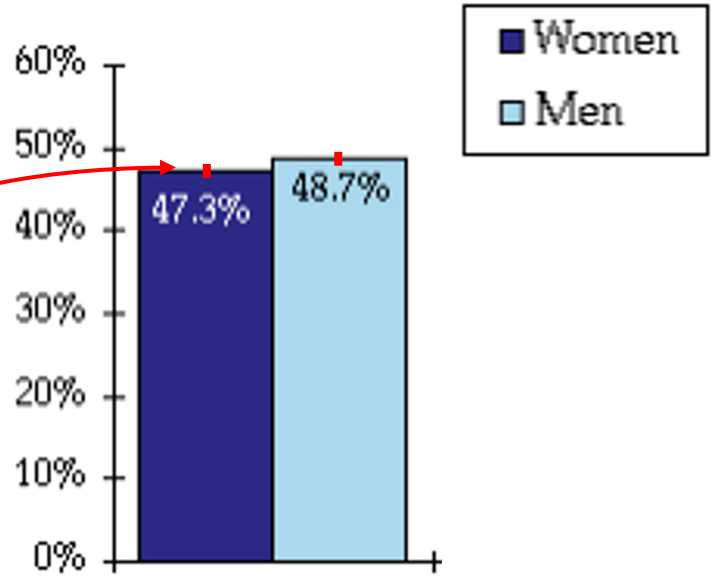
Who likes Snickers better?





Who likes Snickers better?





Be sure to quantify your uncertainty!

- Even a complete dataset is a sample!
- Whenever you report a statistic, you need to quantify how certain you are in it!
- We will discuss two ways of quantifying uncertainty:
  (1) Hypothesis testing
  (2) Confidence intervals
- All plots should have error bars!

How to quantify uncertainty?
Approach 1:

Hypothesis testing

Hypothesis testing: intro

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?

Hypothesis testing: intro

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…

Hypothesis testing

- A huge subject; can take entire classes on it
- A black art; many people hate it
- Need to understand basics even if you don’t use it yourself
- Never use it without understanding exactly what you’re doing

The logic of hypothesis testing

- 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 to explain the data (and is thus not rejected)
- Otherwise, keep experimenting

The logic of 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 that is likely to be **large under H_A** but **small under H_0**

Coin example

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

- **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.

Another example

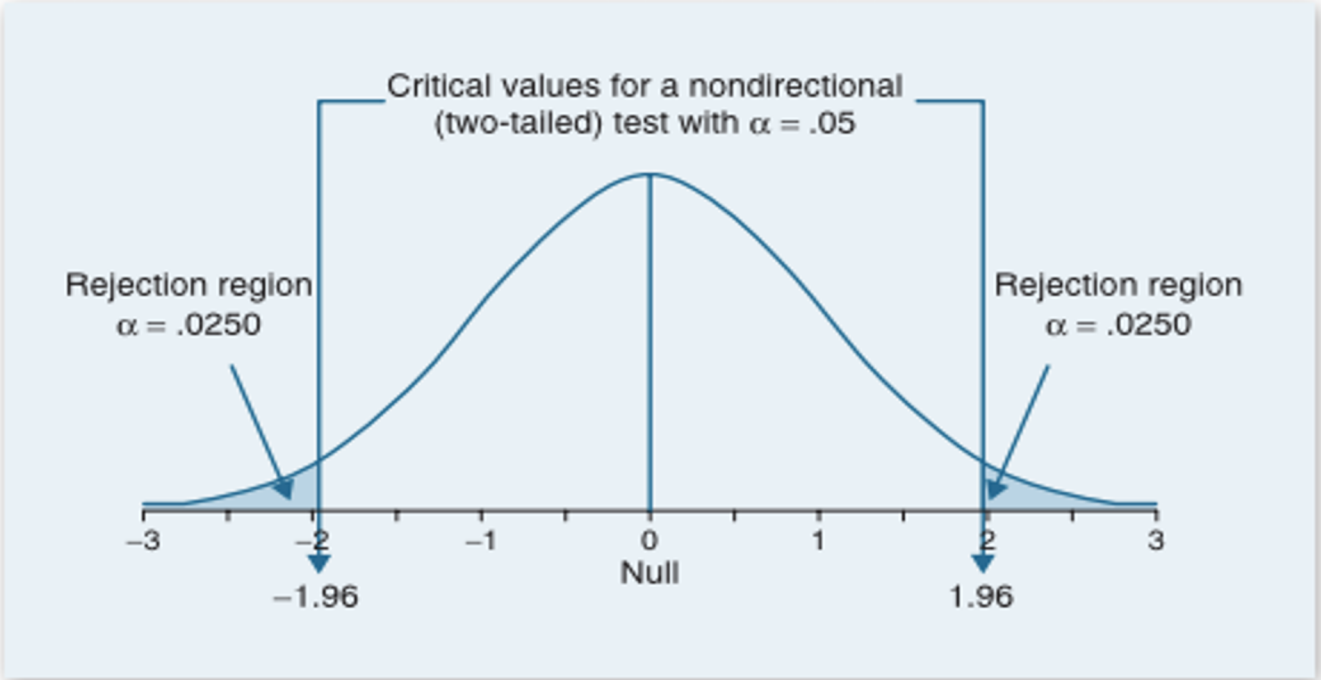
- Two samples *a* and *b*, normally distributed, from *A* and *B*.
- Null hypothesis H_0: mean(*A*) = mean(*B*)
  test statistic is: s = mean(*a*) – mean(*b*).
- Under H_0, *s* has mean zero and is normally distributed*.
- But it is “large” if the two means are different.

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

Another example (cont'd)

- 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 also called the **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



Test statistic in blue tails → (two-sided) p < 0.05 → reject the null hypothesis

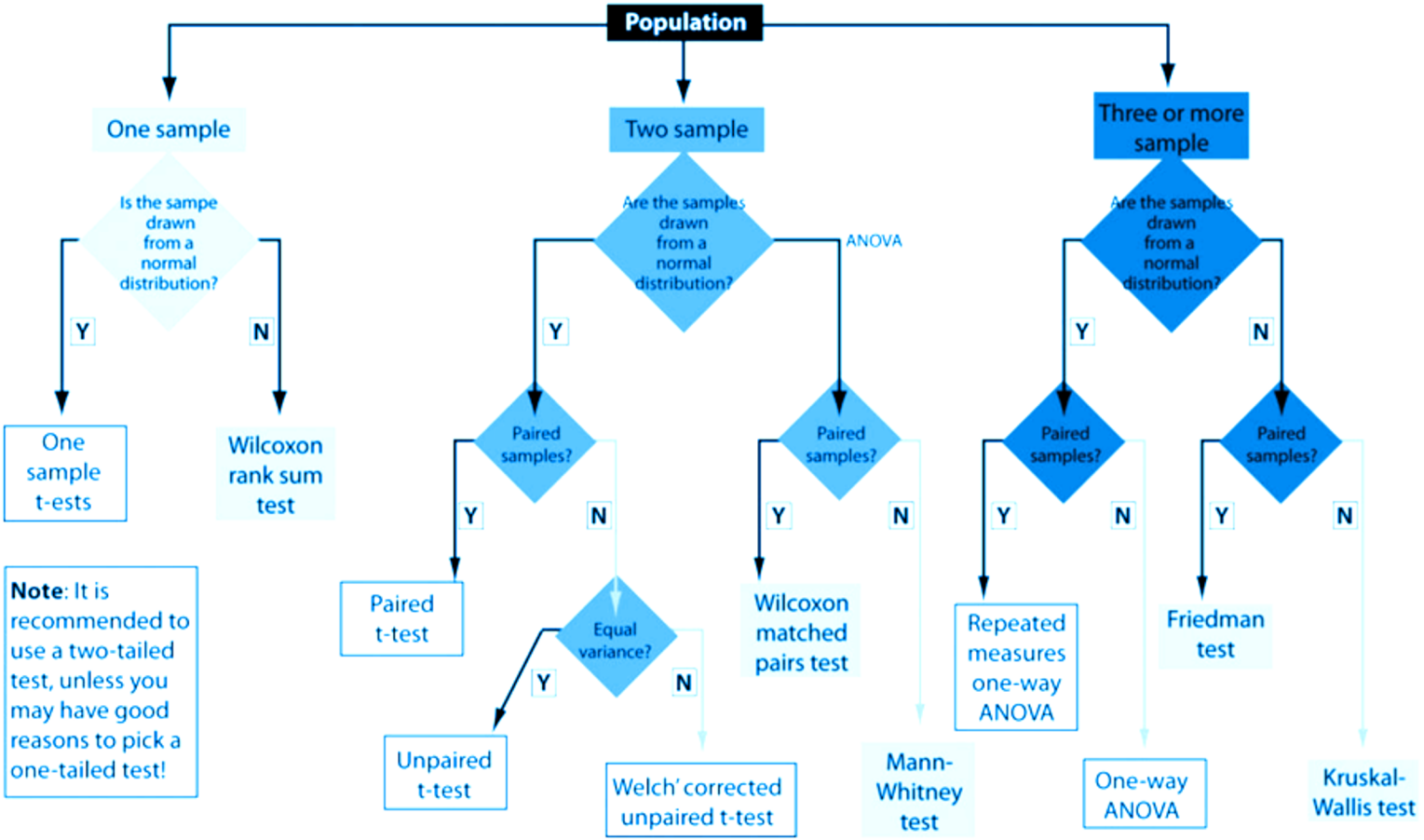
Selecting a test

There are many statistical tests.
Although they differ in their details, the basic logic is always the same (previous slides)

The right choice of test depends on multiple factors:

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

Good news: **Plenty of advice available (p.t.o.)**


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!

Remarks on 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

Remarks on 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"

Remarks on 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

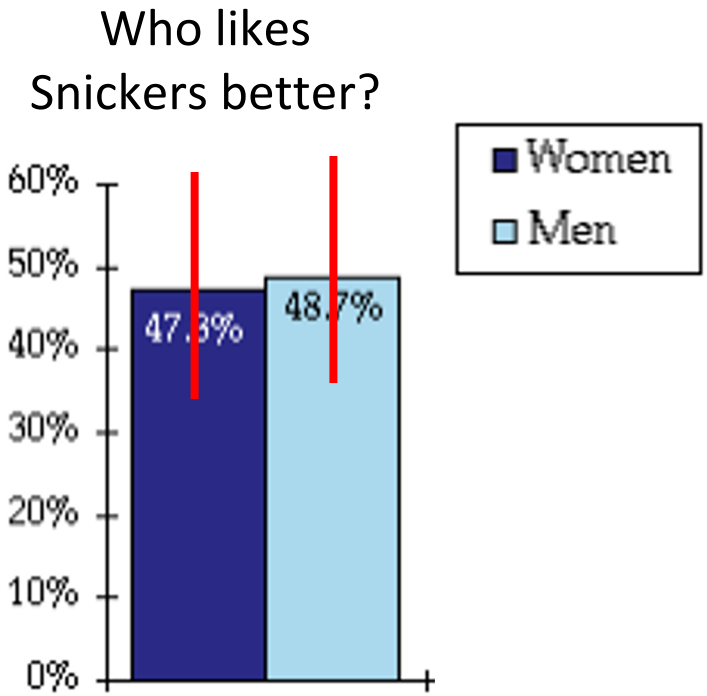
1	*If P = .05, the null hypothesis has only a 5% chance of being true.*
2	*A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.*
3	*A statistically significant finding is clinically important.*
4	*Studies with P values on opposite sides of .05 are conflicting.*
5	*Studies 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.*

How to quantify uncertainty?
Approach 2:

Confidence intervals

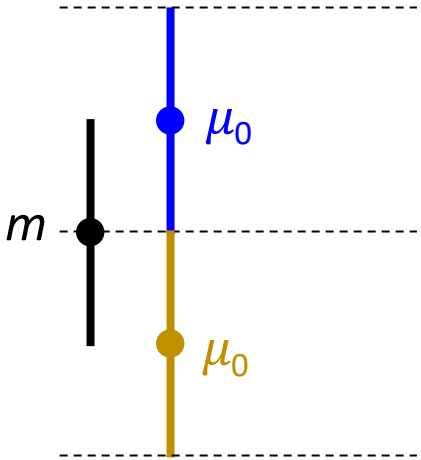
Confidence intervals: idea

- Confidence interval (CI)
  = a range of estimates for the parameter of interest (e.g., mean) that seems reasonable given the observed data
- Confidence level γ ⇒ "γ CI"
  (usually γ = 95% ⇒ 95% CI)



Confidence intervals: definition

- μ: true value of parameter of interest
- m: empirical estimate of parameter of interest
- CIs and hypothesis testing are tightly connected:
  - γ CI contains those values μ_0 for which the null hypothesis "H_0: $\mu = \mu_0$" cannot be rejected at significance level 1−γ

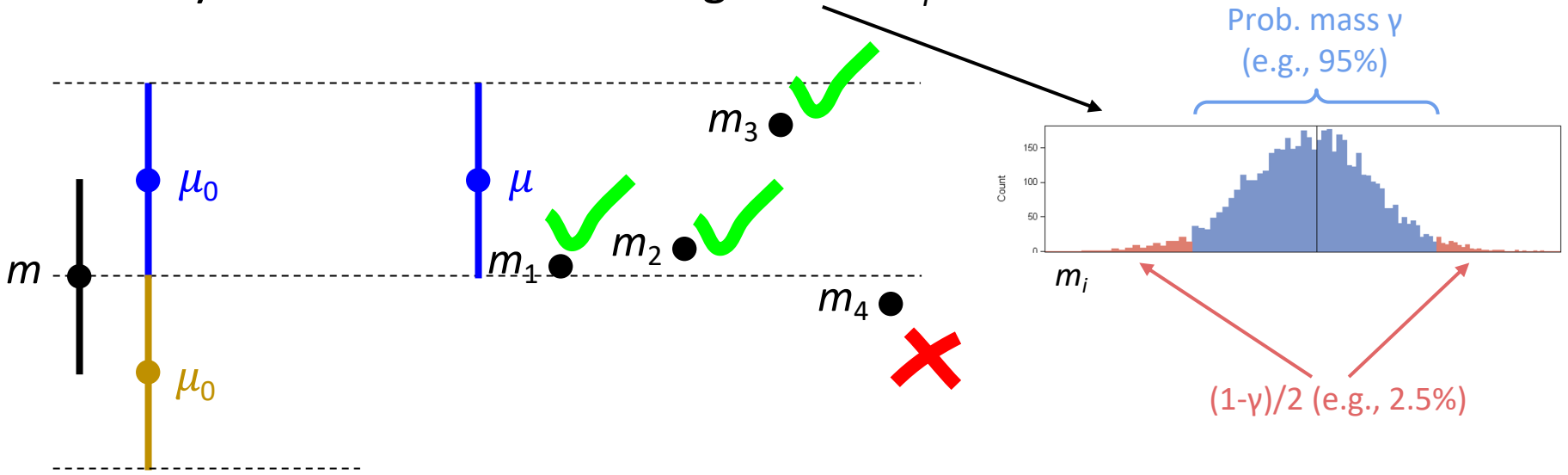


How to compute confidence intervals?

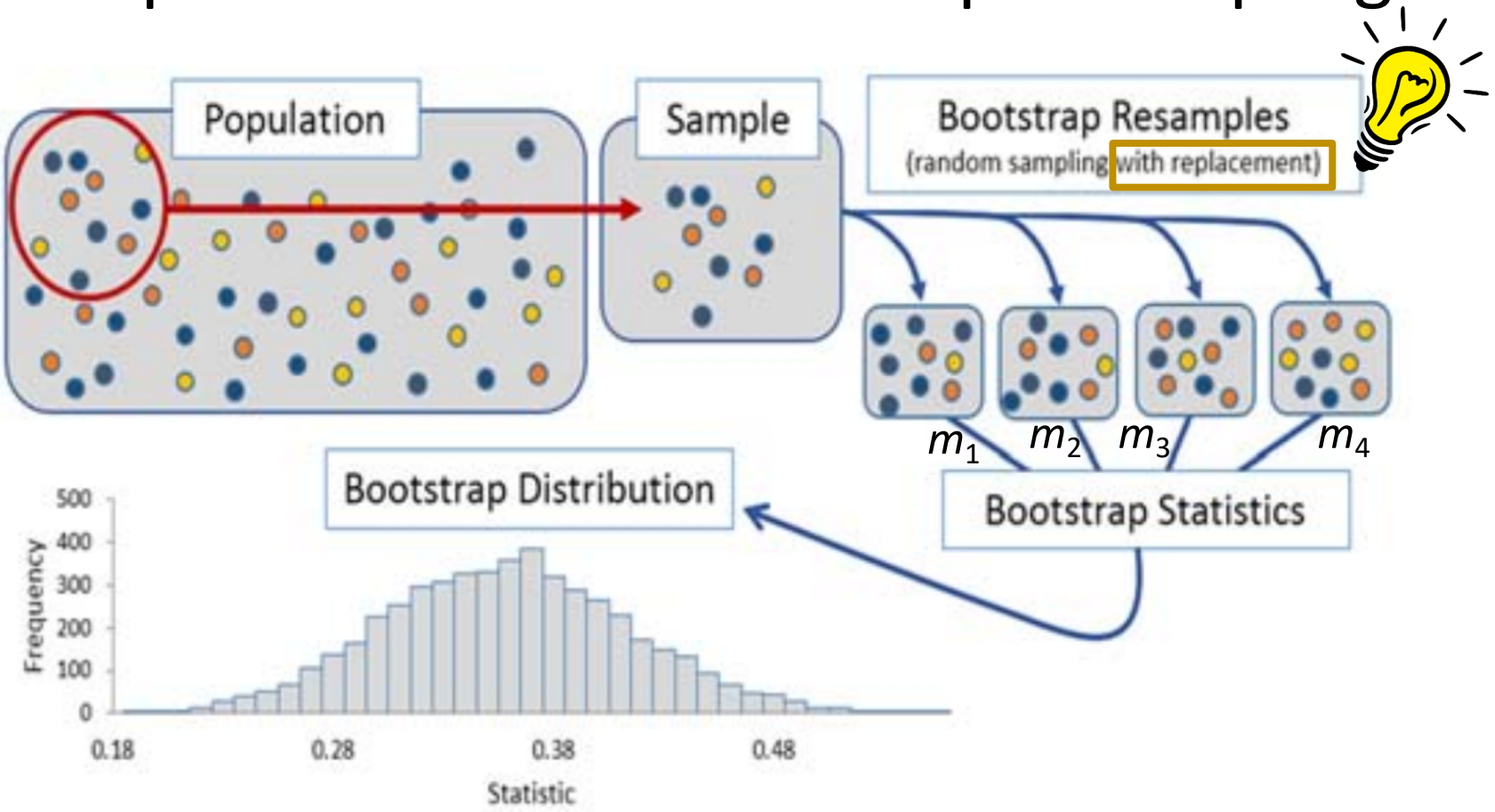
- **Parametric** methods assume that the test statistic follows a known (typically Normal) distribution
  → Need to verify that this is actually true!
- **Non-parametric** methods make no assumptions about the distribution of the test statistic. They instead work by sampling the empirical data.

Confidence intervals: another view

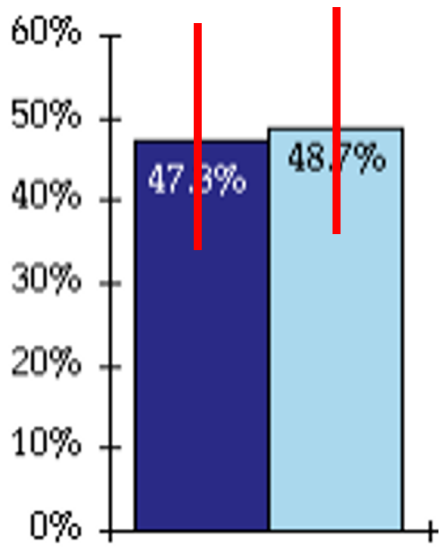
- If we were to repeat the data collection $N \to \infty$ independent times, we'd obtain N estimates of μ: $m_1, \ldots, m_N$
- Average of m_i's (as well as m) will approach the true μ
- For a fraction γ of the N repetitions, m_i lies within the γ CI around μ
- → May estimate CI from histogram of m_i's



Non-parametric CIs: bootstrap resampling



Error bars



- An important use case for CIs
- But be careful! CIs can potentially represent many things:
  - Confidence intervals
  - Standard deviation
  - Standard error of the mean: std/sqrt(n)
- → Always ask, always tell what the CIs represent!

Multiple-hypothesis testing

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

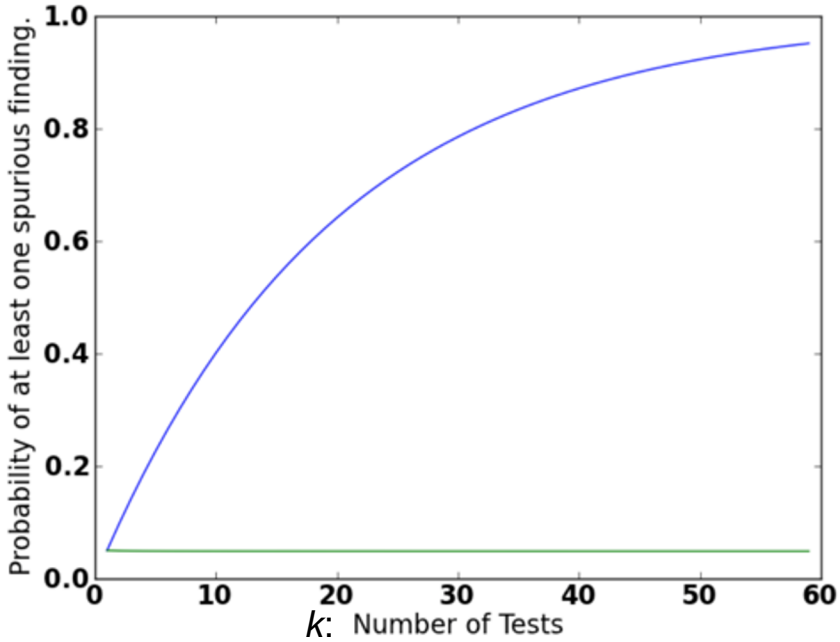
Significance level must be adjusted down when performing multiple hypothesis tests!

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



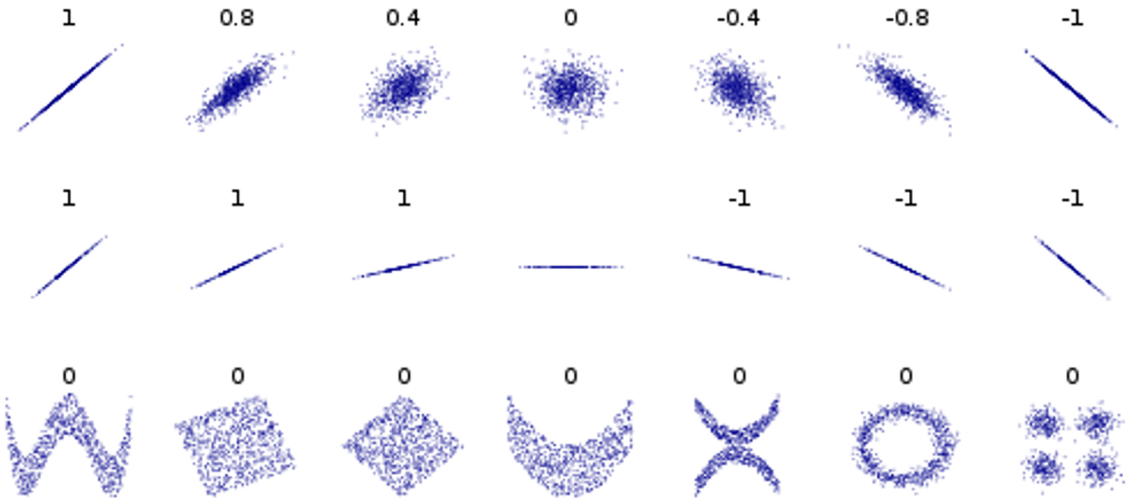
$\alpha = 0.05$

"Familywise Error Rate"

Part 3
Relating two variables

Pearson's correlation coefficient

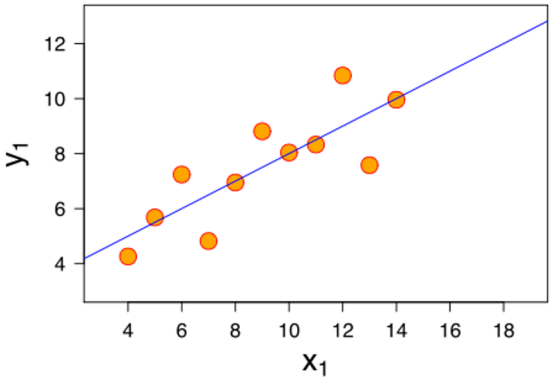
- "Amount of linear dependence"

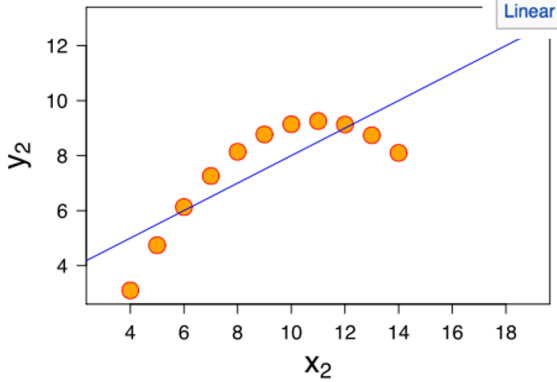


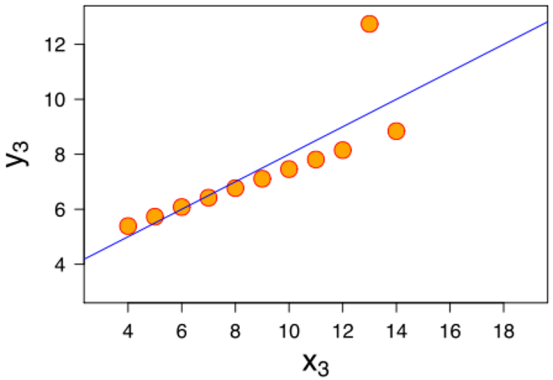
- More general:
  - Rank correlation, e.g., Spearman's correlation coefficient
  - Mutual information

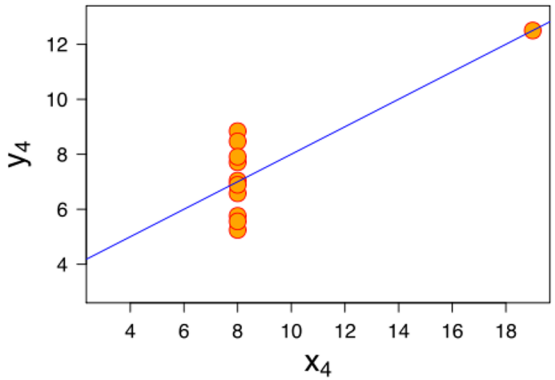
Anscombe's quartet

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

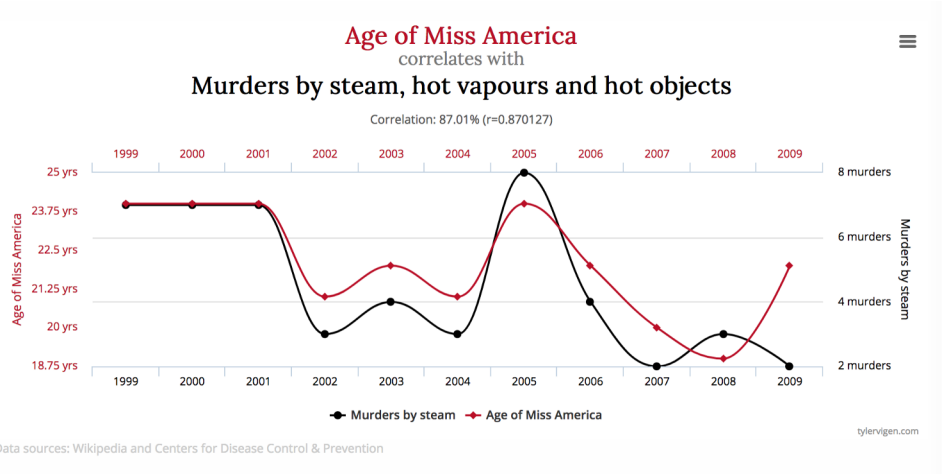
Highlights the *inadequacy of basic statistical properties for describing realistic datasets*

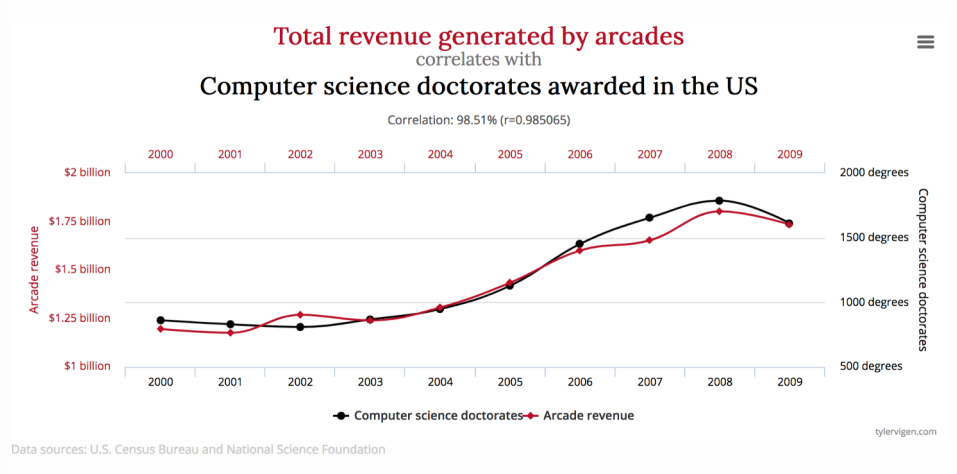
More on Wikipedia

Correlation coefficients are tricky!

- http://guessthecorrelation.com/
- Correlation != causation

  http://www.tylervigen.com/spurious-correlations



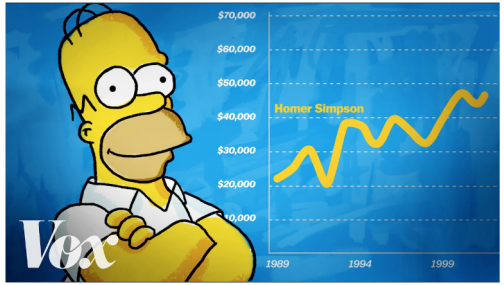


UC Berkeley gender bias (?)

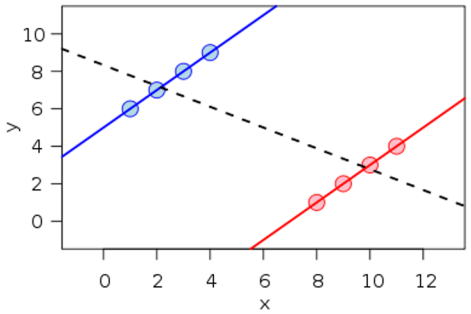
Admission figures from 1973

	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

Summary

- Understand your data with descriptive statistics
  - Choose the right stats based on type of distribution
- Be sure to quantify your uncertainty
  - Hypothesis testing
  - Confidence intervals (preferred!)
  - Careful when performing multiple tests (apply correction)
- Relating 2 variables to one another
  - Correlation != causation

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

Benjamin Disraeli (1804 - 1881)