Applied Data Analysis (CS401)

Lecture 4
Describing data
11 Oct 2023

EPFL

Robert West

dlab

Announcements

- Project milestone P1 due this Fri 13 Oct 23:59
  - Remember: we won't answer questions in final 24h
- Homework H1 to be released Fri 13 Oct, due Fri 27 Oct
- Friday's lab session:
  - From now on: one single room: BCH2201
  - Exercises on topic of this lecture (describing data)
  - Quiz 3
- Next week's Wed lecture: held on Zoom due to travel
  - You can watch a live stream in Rolex Learning Center or watch from home

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec4-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? / Où est Charlie?
- …

Overview of today's lecture

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

ADA won't cover the basics of stats!

You know these things from prerequisite courses

But stats is a key ingredient of data analysis

Today: some highlights and common pitfalls

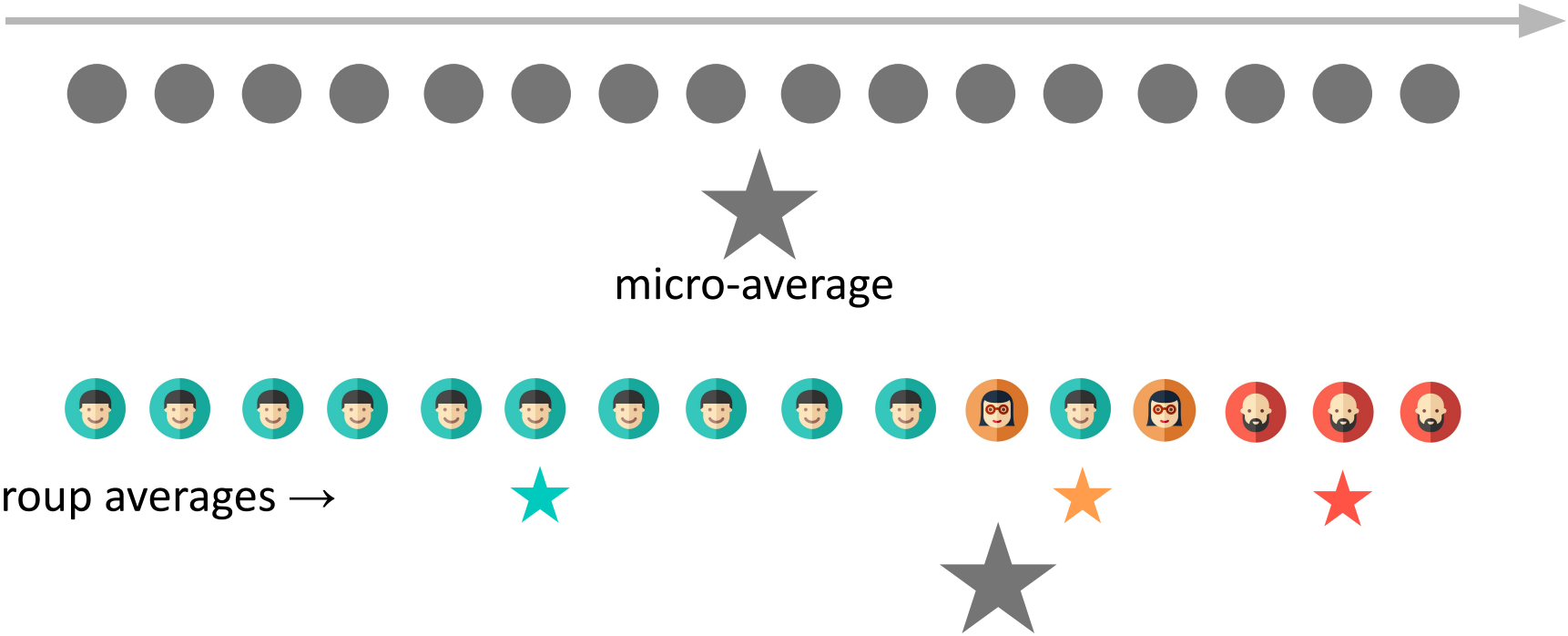
Part 1
Descriptive statistics

Descriptive statistics

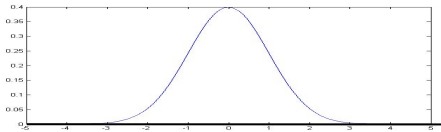
```
baseball.describe()
```

	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

Means: micro- vs. macro-average



Robust statistics



x

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

```
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

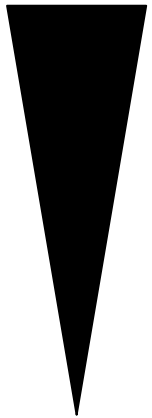
- Some distributions are all about the extreme values
- E.g., power laws (see last lecture): $f(x) = ax^{-k}$
  - Very very large values are rare, “but not very rare”
  - Body size vs. city size
  - For k <= 3: infinite variance
  - For k <= 2: infinite variance, infinite mean
  - Don’t report (arithmetic)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

Generalized means [Wikipedia]

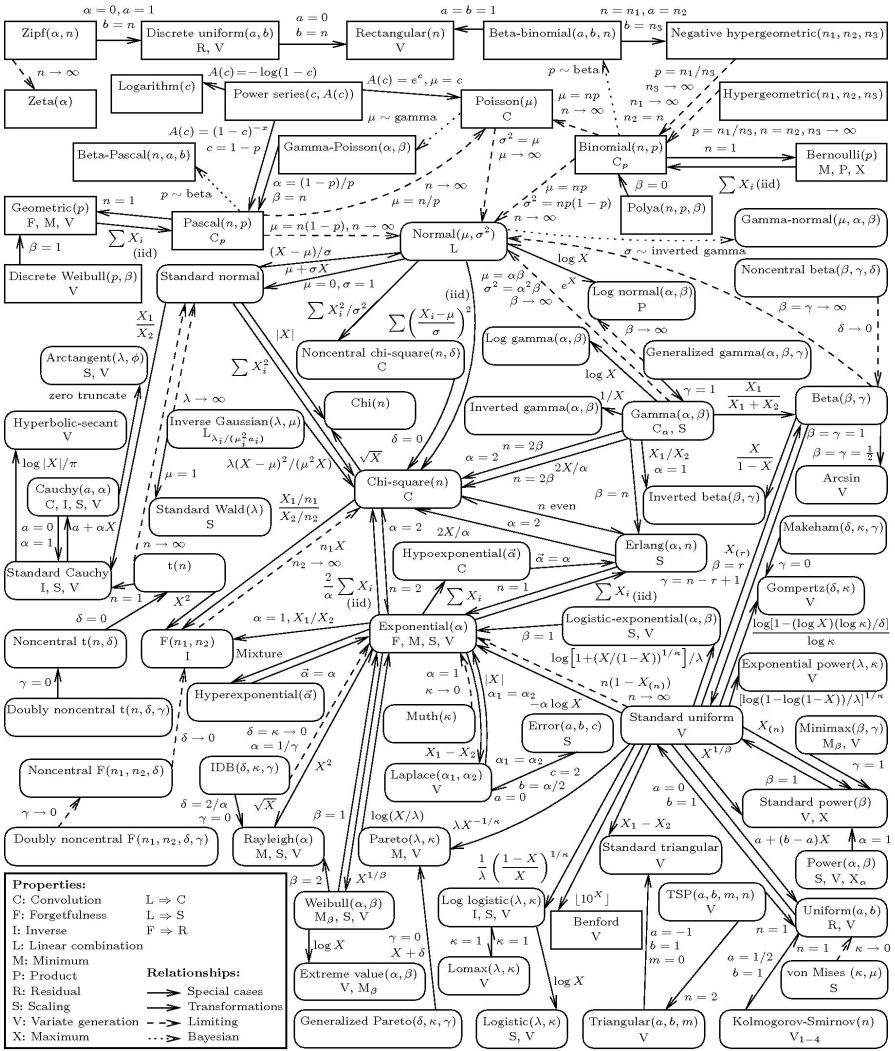
- Common trick: transform data into a different space (via function f), take mean there, then transform back into the original space (via f^{-1}):

$$f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right)$$

- $f(x) = x^2$, $f^{-1}(x) = \sqrt{x}$ “root mean square”
- $f(x) = x$, $f^{-1}(x) = x$ “arithmetic mean”
- $f(x) = \log(x)$, $f^{-1}(x) = \exp(x)$ “geometric mean”
- $f(x) = 1/x$, $f^{-1}(x) = 1/x$ “harmonic mean”



Distributions



[link]

Distributions

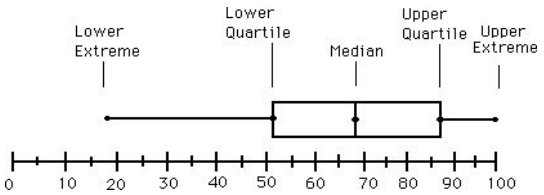
Some important distributions:

- **Normal:** see previous slides
- **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 “successes” (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

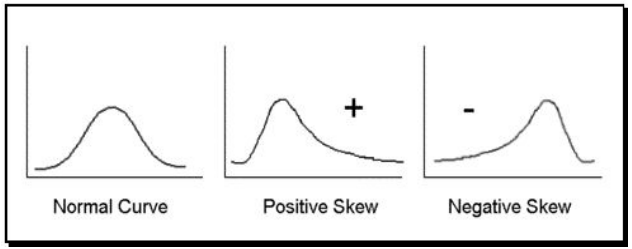
You should understand the distribution of your data before applying any model!

“Dear data, where are you from?”

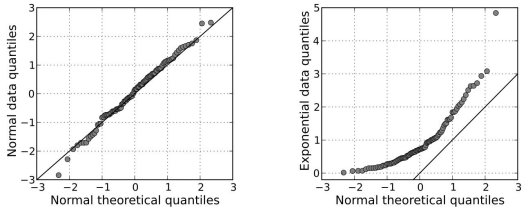
- Visual inspection for ruling out certain distributions:
  e.g., when histogram/box plot is asymmetric (even for large sample size), the data cannot be normal



Box plot



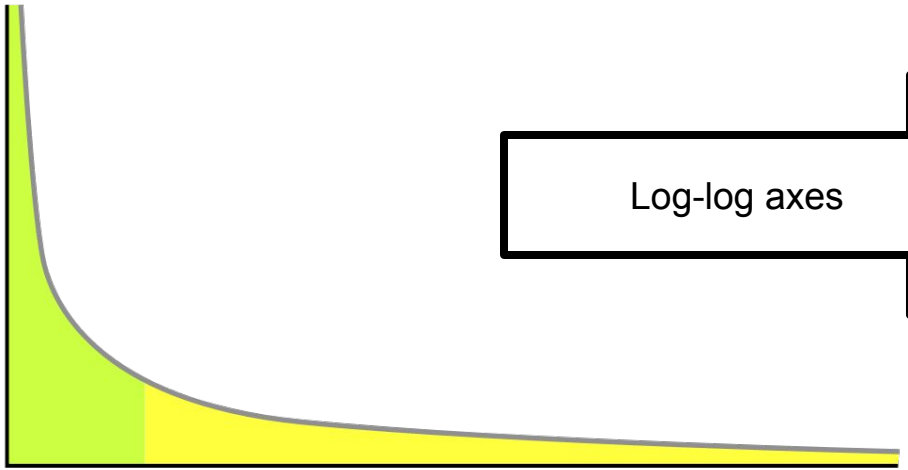
(Smoothed) histogram



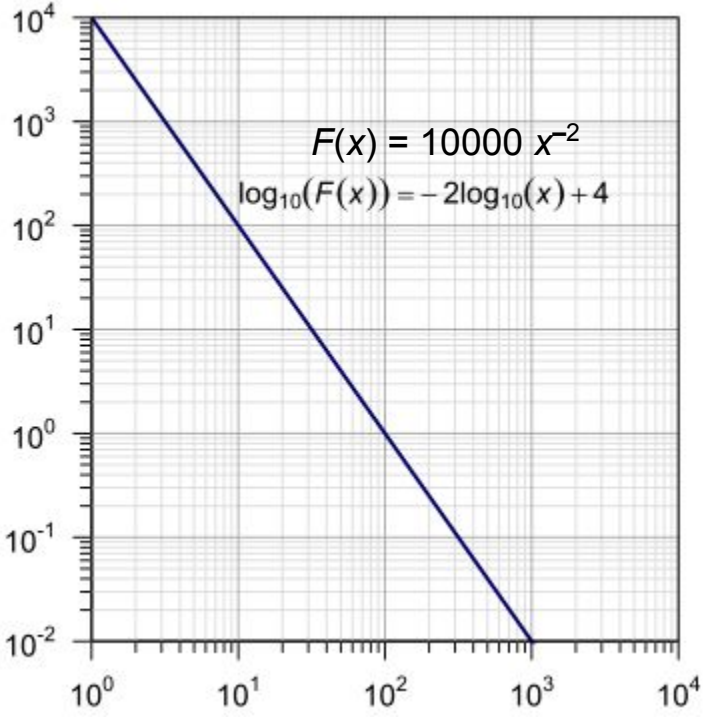
Quantile-quantile (QQ) plots

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

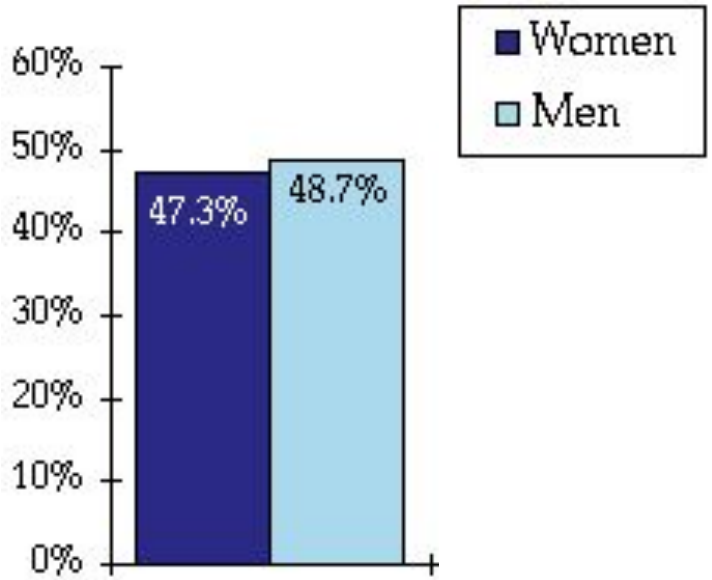
Recognizing a power law



Log-log axes



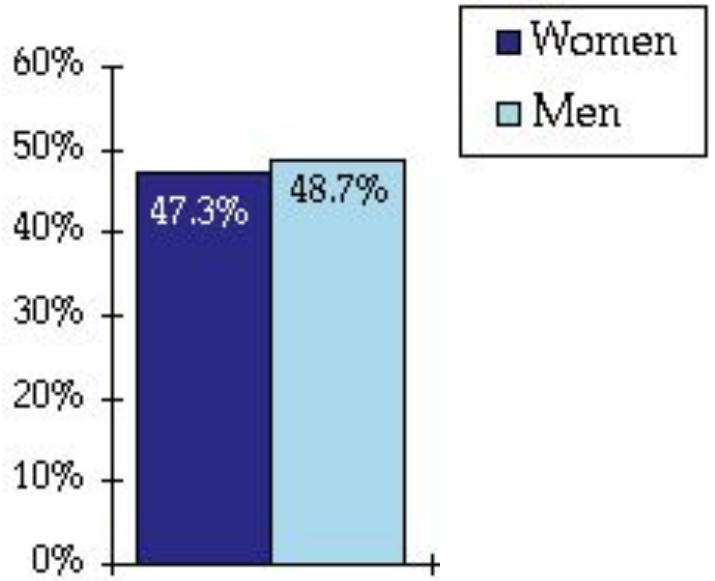
Who likes Snickers better?



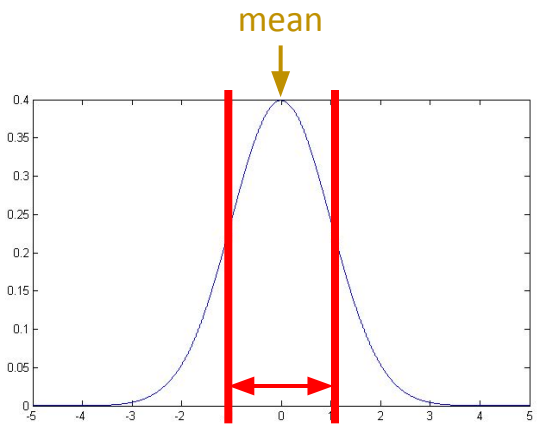
Part 2
Quantifying uncertainty

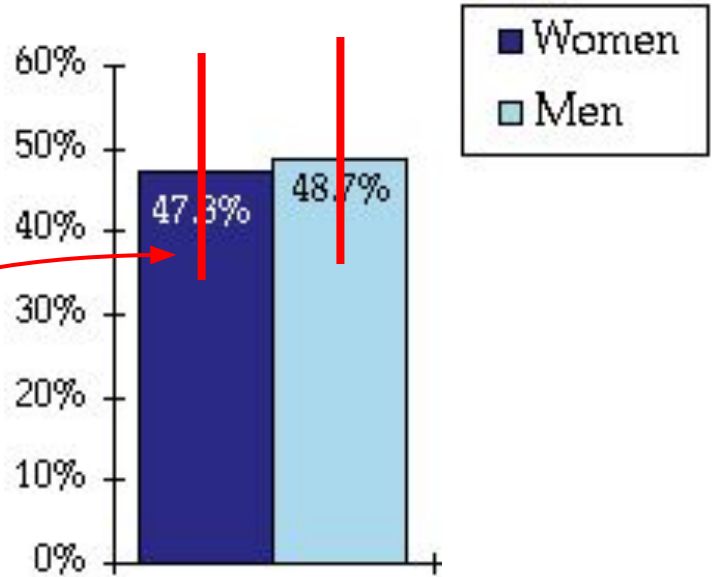
Who likes Snickers better?

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

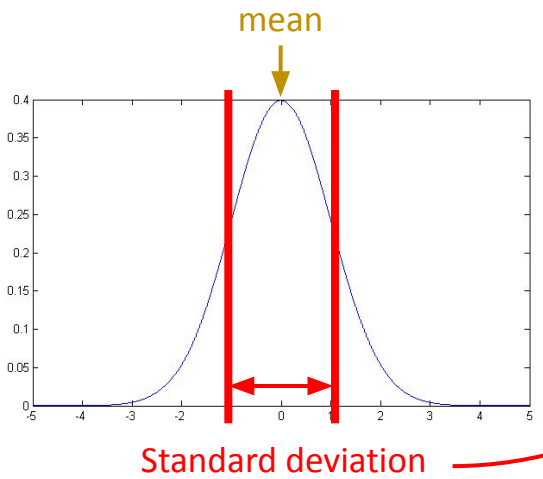


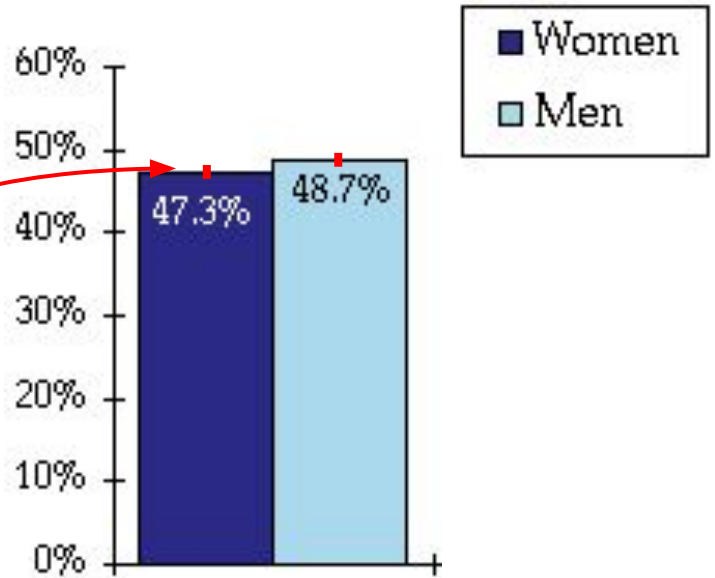
Who likes Snickers better?





Who likes Snickers better?





Be certain to quantify your uncertainty!

- Finite samples introduce uncertainty
  - Even a complete dataset is a finite 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

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.*

THINK FOR A MINUTE:

Which of these statements about p-values are true?

(Feel free to discuss with your neighbor.)

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.*

POLLING TIME

- **"Which of these statements about p-values are true?"**
- Scan QR code or go to https://web.speakup.info/room/join/66626

Hypothesis testing: intro

Joseph B. 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
- Many people don't like 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

Commercial break

The logic of hypothesis testing

- Flip a coin 100 times; outcome: 40 heads; “Is the coin fair?”
- Null hypothesis: “yes”; alternative hypothesis: “no”
- “How likely would I be to see an outcome at least this extreme (i.e., ≤ 40 heads) if the null hypothesis were true (i.e., if the coin were fair, i.e., if we expect 50 heads)?”
- If this probability is large, the null hypothesis suffices to explain the data (and is thus not rejected)
- Otherwise, dig deeper in order to understand your data

The logic of hypothesis testing

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

- by inspecting a **test-statistic:** a measurement made on the data that is likely to be large under H_A but small under H_0

Coin example

- Idea: Gain (weak and indirect) support for a hypothesis H_A by **ruling out a null hypothesis H_0**
  - H_0: “the coin is fair” (simplest hypothesis, cf. Occam’s razor)
  - H_A: “the coin is not fair (a.k.a. biased)”

- by inspecting a **test-statistic:** a measurement made on the data that is likely to be large under H_A but small under H_0
  - e.g., number of heads after 100 coin tosses (1-tailed)
  - e.g., abs(50 - number of heads after 100 coin tosses) (2-tailed)

Coin example (cont'd)

- **Null hypothesis H_0:** "the coin is fair", i.e., "probability of heads = 0.5"
- **Test statistic s:** abs(50 - # of empirically observed heads after 100 coin tosses)
- Pr(S | H_0): probability distribution of test statistic, assuming that H_0 is true

- Decision rule: **reject H_0 if Pr(S ≥ s | H_0) < α,**
  i.e., if the probability of deviating from 50 heads at least as much as empirically observed is small
  - Pr(S ≥ s | H_0) = **"p-value"**
  - α = **"significance level"**
- α controls "false-rejection rate" (probability of rejecting H_0 although it is true)
  - You as the data analyst choose α (common values: 5%, 1%, 0.5%, 0.1%)
  - Higher α → higher false-rejection rate

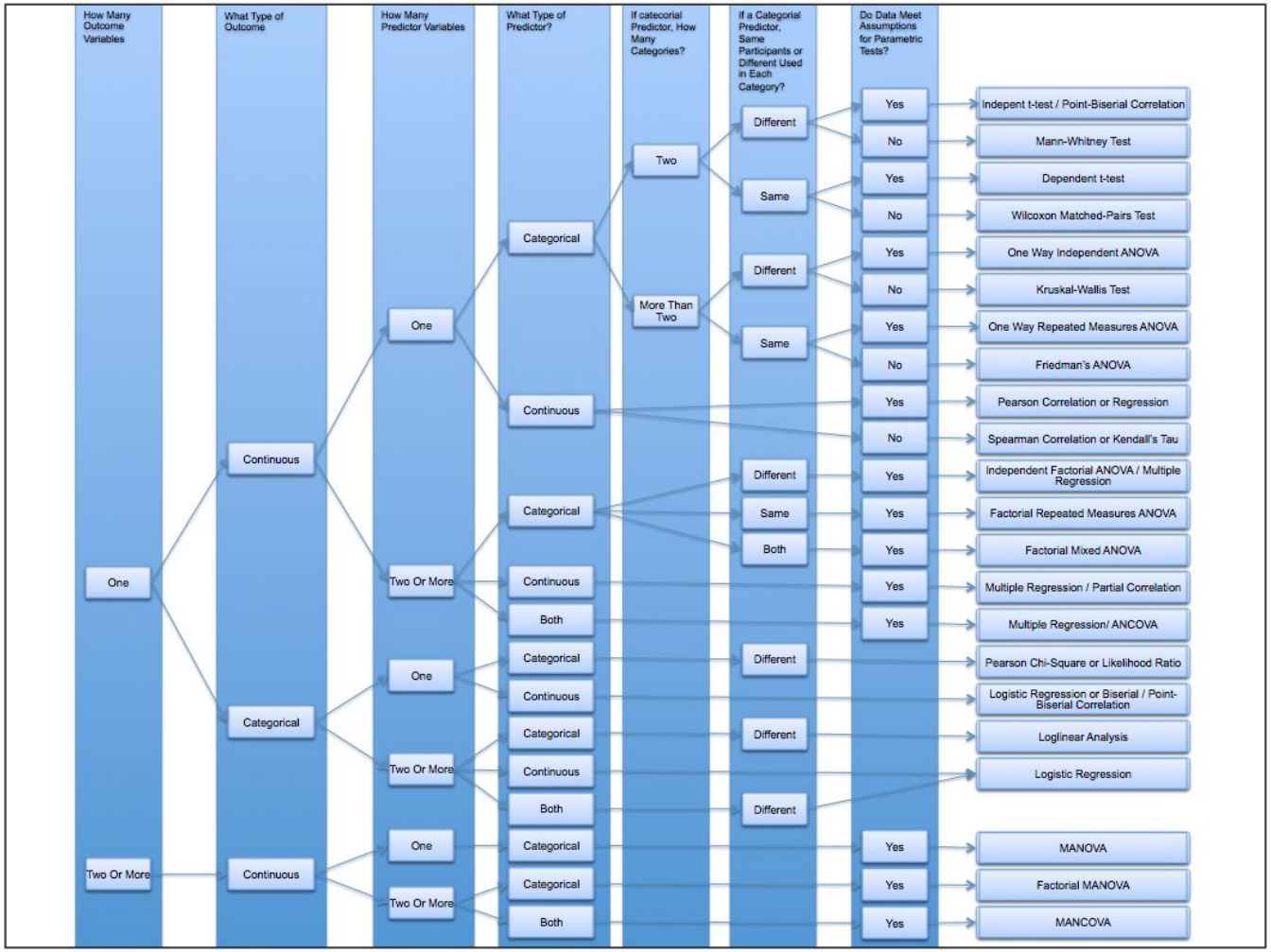
Selecting the right test

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

The right choice of test depends on multiple factors (here a selection):

- Question
- Data type (continuous vs. categorical; dimensionality; number of outcomes)
- Sample size
- When comparing two samples: same population or different populations?
- Parametric assumptions about distribution of test statistic under null hypothesis? (Less important for large sample sizes, due to central limit theorem)

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


How Many Outcome Variables
What Type of Outcome
How Many Predictor Variables
What Type of Predictor?
If catecorial Predictor, How Many Categories?
If a Categorial Predictor, Same Participants or Different Used in Each Category?
Do Data Meet Assumptions for Parametric Tests?
One
Two Or More
Continuous
Categorical
One
Two Or More
One
Two Or More
Categorical
Continuous
Categorical
Continuous
Both
Two
More Than Two
Categorical
Continuous
Both
Categorical
Continuous
Both
Different
Same
Different
Same
Different
Same
Both
Different
Different
Different
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Indepent t-test / Point-Biserial Correlation
Mann-Whitney Test
Dependent t-test
Wilcoxon Matched-Pairs Test
One Way Independent ANOVA
Kruskal-Wallis Test
One Way Repeated Measures ANOVA
Friedman's ANOVA
Pearson Correlation or Regression
Spearman Correlation or Kendall's Tau
Independent Factorial ANOVA / Multiple Regression
Factorial Repeated Measures ANOVA
Factorial Mixed ANOVA
Multiple Regression / Partial Correlation
Multiple Regression/ ANCOVA
Pearson Chi-Square or Likelihood Ratio
Logistic Regression or Biserial / Point-Biserial Correlation
Loglinear Analysis
Logistic Regression
MANOVA
Factorial MANOVA
MANCOVA

Remarks on p-values

- Widely used in all sciences
- They are widely misunderstood!
- Don’t use them if you don’t understand them!
- Large p means that even under a simple null hypothesis your data would be quite likely
- This tells you nothing about the alternative 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”
- Look at the effect size (“y-axis”), not just the p-value!

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.*

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

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

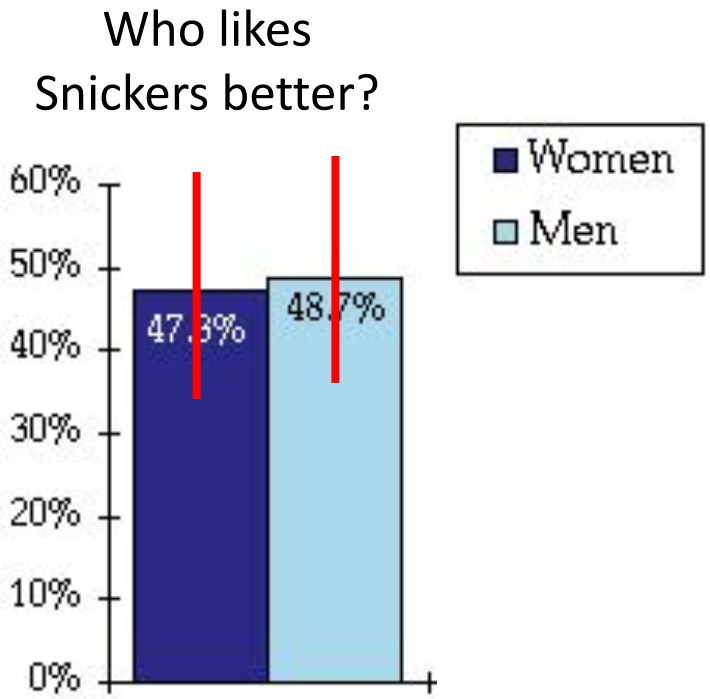
- See here
- Great (and amusing) explanation of difference between hypothesis-testing approach and Bayesian approach: Chapter 37 in MacKay's (free) book on "Information Theory, Inference, and Learning Algorithms"

How to quantify uncertainty?
Approach 2:

Confidence intervals

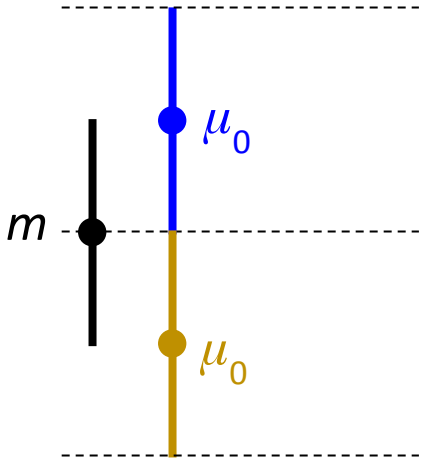
Confidence intervals: idea

- 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"
  (often 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



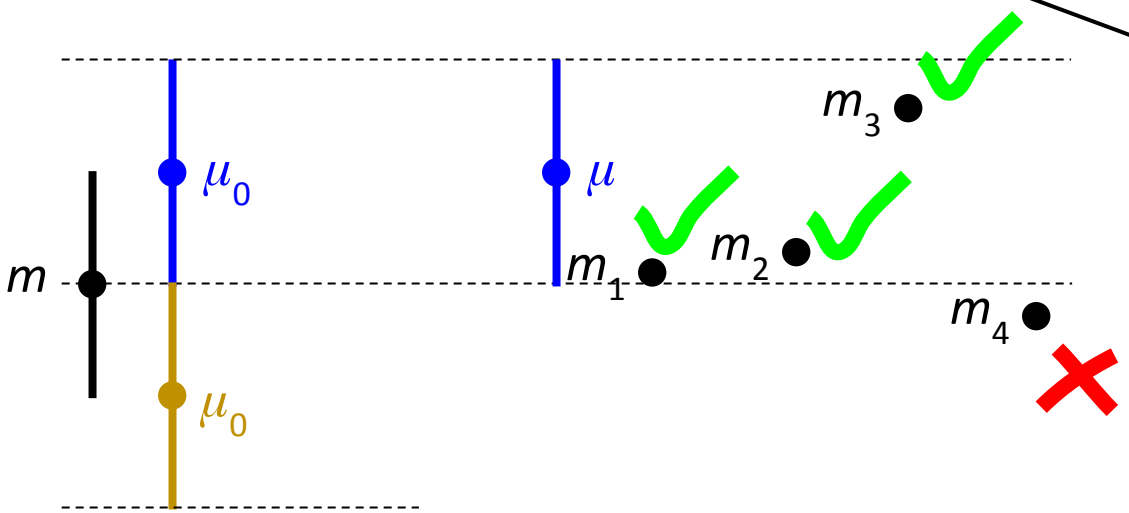
How to compute confidence intervals?

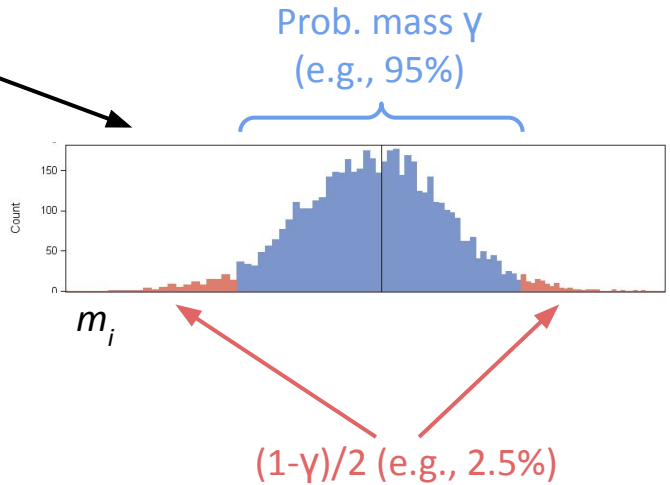
- **Parametric** methods assume that the test statistic follows a known (typically Normal) distribution
  → Need to verify that this is actually true! Ugh…

- **Non-parametric** methods make no assumptions about the distribution of the test statistic. They instead work by sampling the empirical data.
  → Yay!

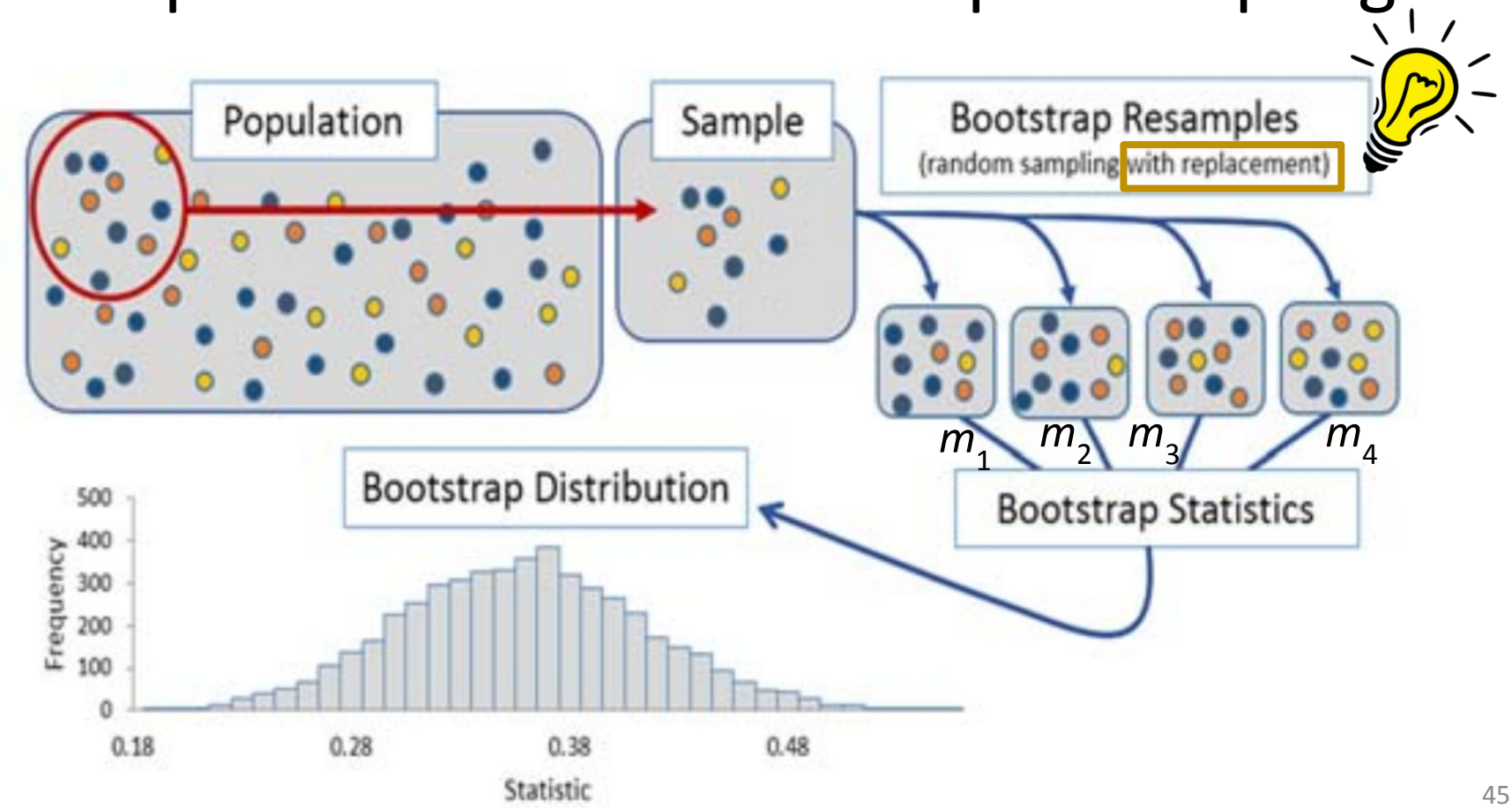
Confidence intervals: another view

- 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 will approach the true μ (by law of large numbers)
- 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

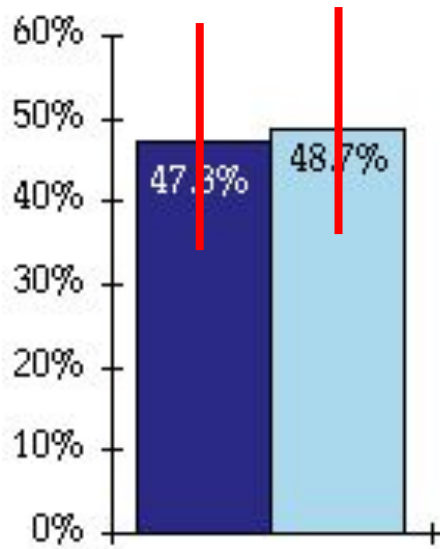




Non-parametric CIs: bootstrap resampling



Error bars



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

Multiple-hypothesis testing

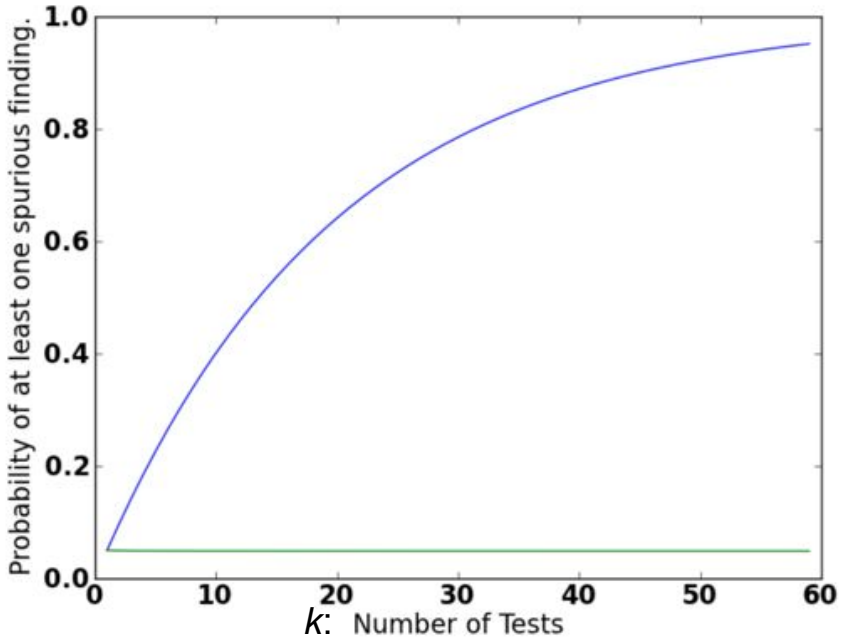
- If you perform experiments over and over, you're bound to find something
- If you consider "at least one positive outcome" to be the manifestation of an underlying effect: Significance level must be adjusted down when performing multiple hypothesis tests!

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

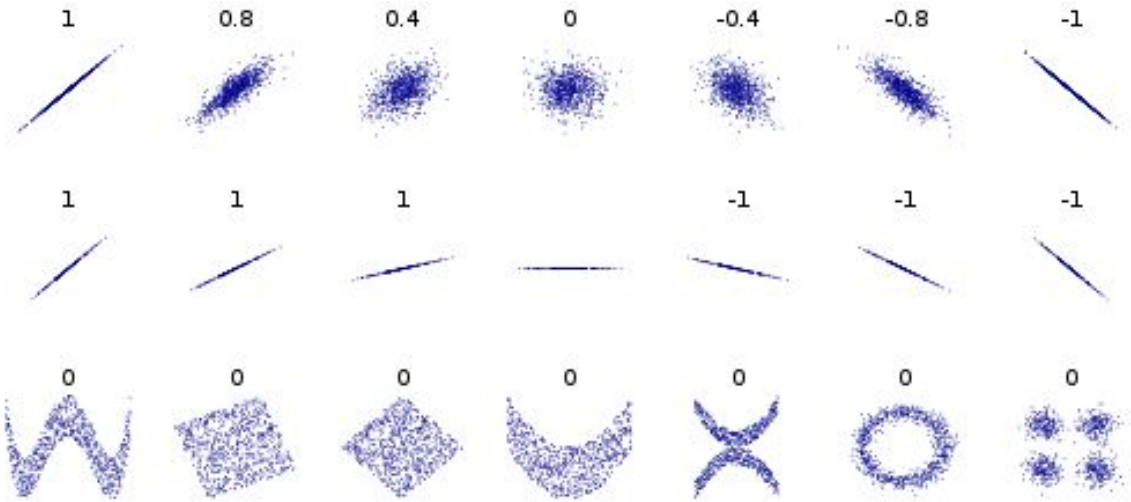
$$\alpha = 1 - (1 - \alpha_c)^k$$

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

Part 3
Relating two variables

Pearson's correlation coefficient

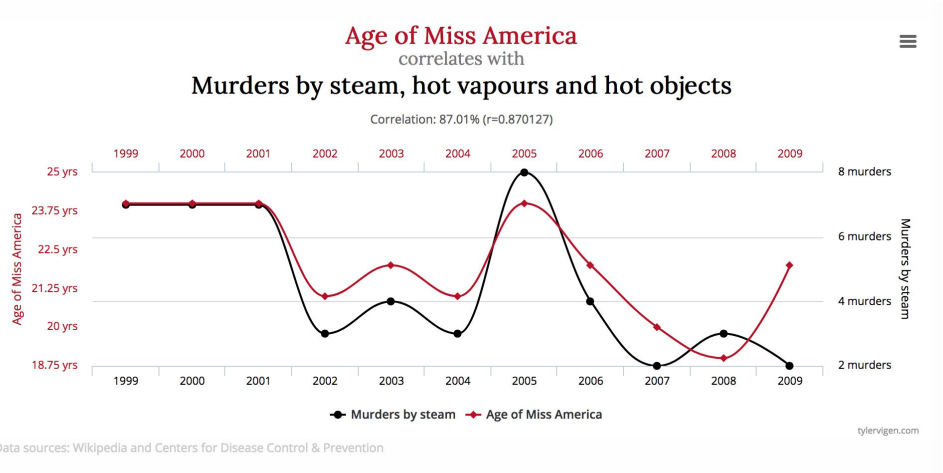
- "Amount of linear dependence"

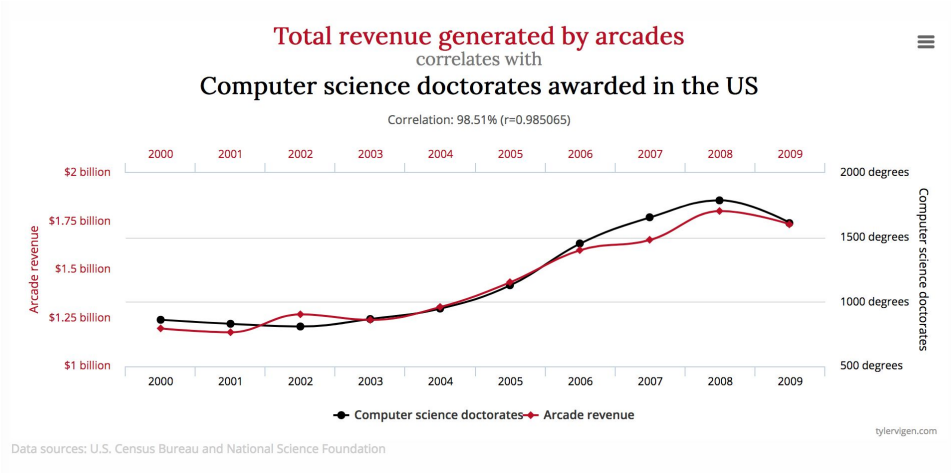


- More general:
  - Rank correlation, e.g., Spearman's correlation coefficient
  - Mutual information

Correlation coefficients are tricky!

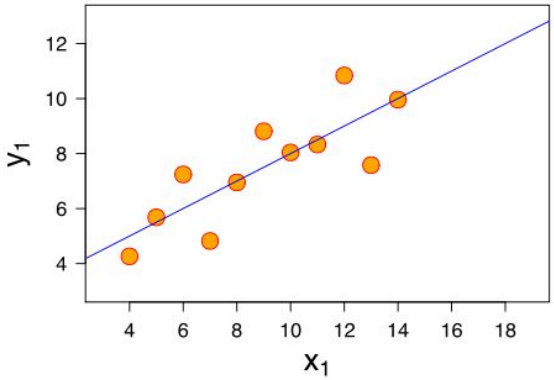
- http://guessthecorrelation.com/
- Correlation != causation (cf. lecture in 2 weeks)
  http://www.tylervigen.com/spurious-correlations

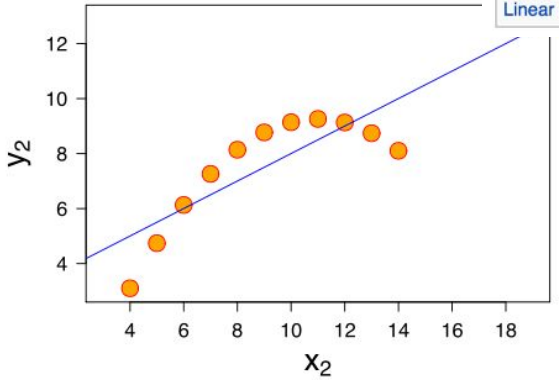


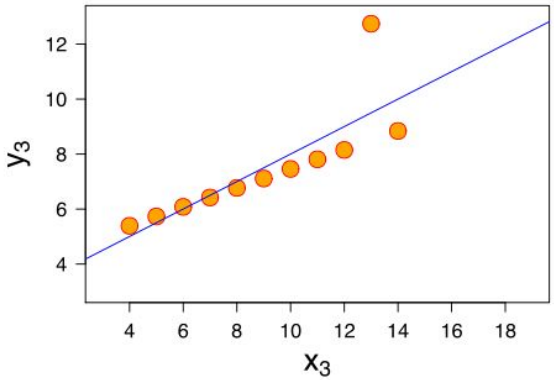


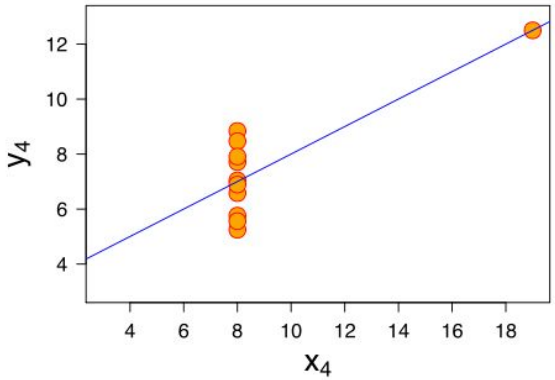
Anscombe's quartet

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

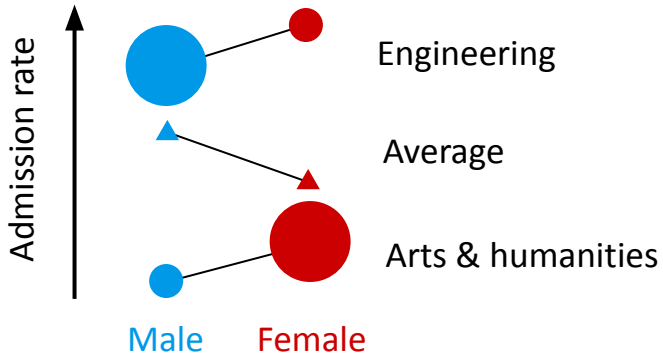
Highlights the *inadequacy of basic statistical properties for describing realistic datasets*

More on Wikipedia

UC Berkeley gender bias (?)

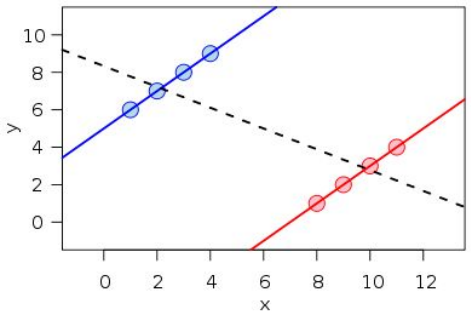
Admission figures from 1973

	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!

In 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 certain to quantify your uncertainty
  - Hypothesis testing
  - Confidence intervals (preferred!)
  - Careful when performing multiple tests (apply correction)
- Relating 2 variables to one another
  - Correlation != causation
  - Even tricker with >2 variables (→ next lecture!)

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec4-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? / Où est Charlie?
- …