Computational Statistics

Today

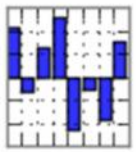
Statistical inference

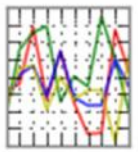
- Effect size.
- Quantifying precision.
- Hypothesis testing.

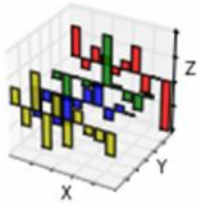
python™

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$







Interesting times

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For A

Archive > Volume 519 > Issue 7541 > Research Highlights: Social Selection > Article

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

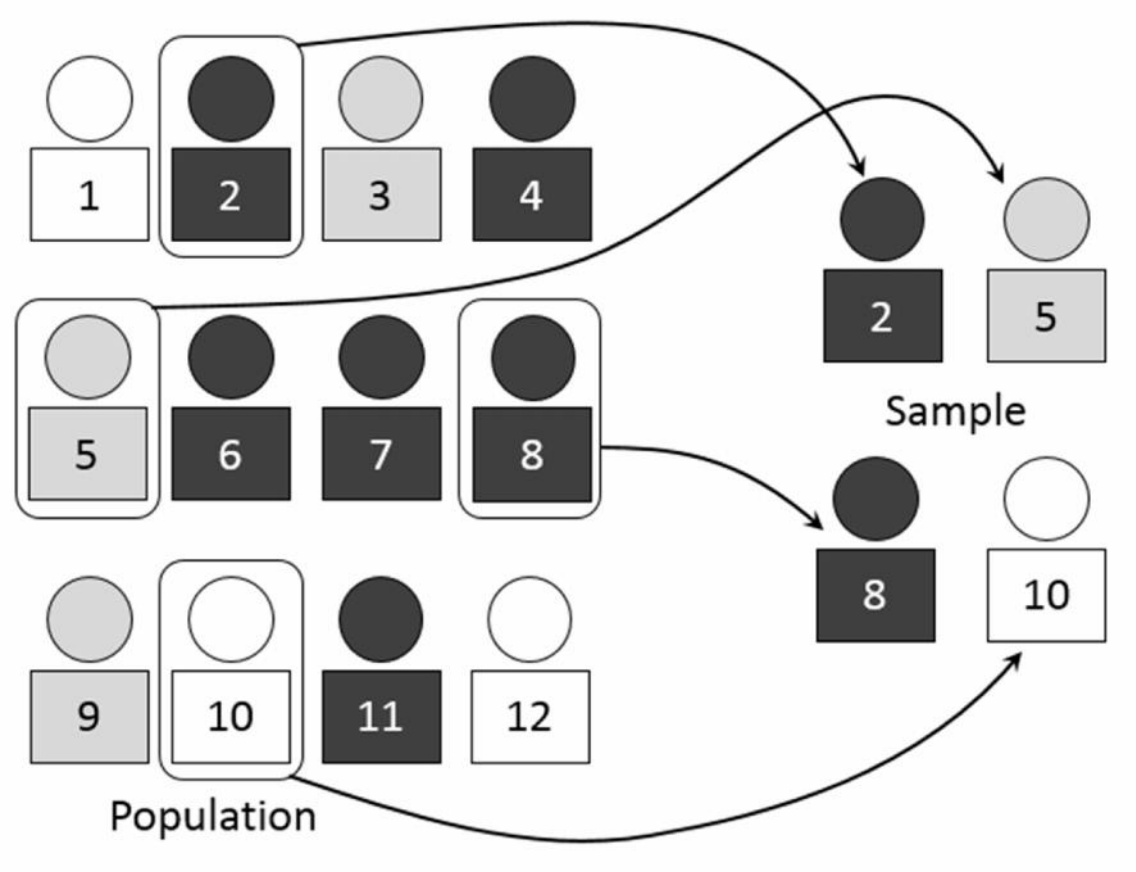
26 February 2015 | Clarified: 09 March 2015

What’s the problem?

What is statistical inference?
What are we doing wrong?

Statistical inference

Using data from a sample to infer information about a population.



Simple random sampling by Dan Kernler via Wikimedia Commons.

Example: drug testing

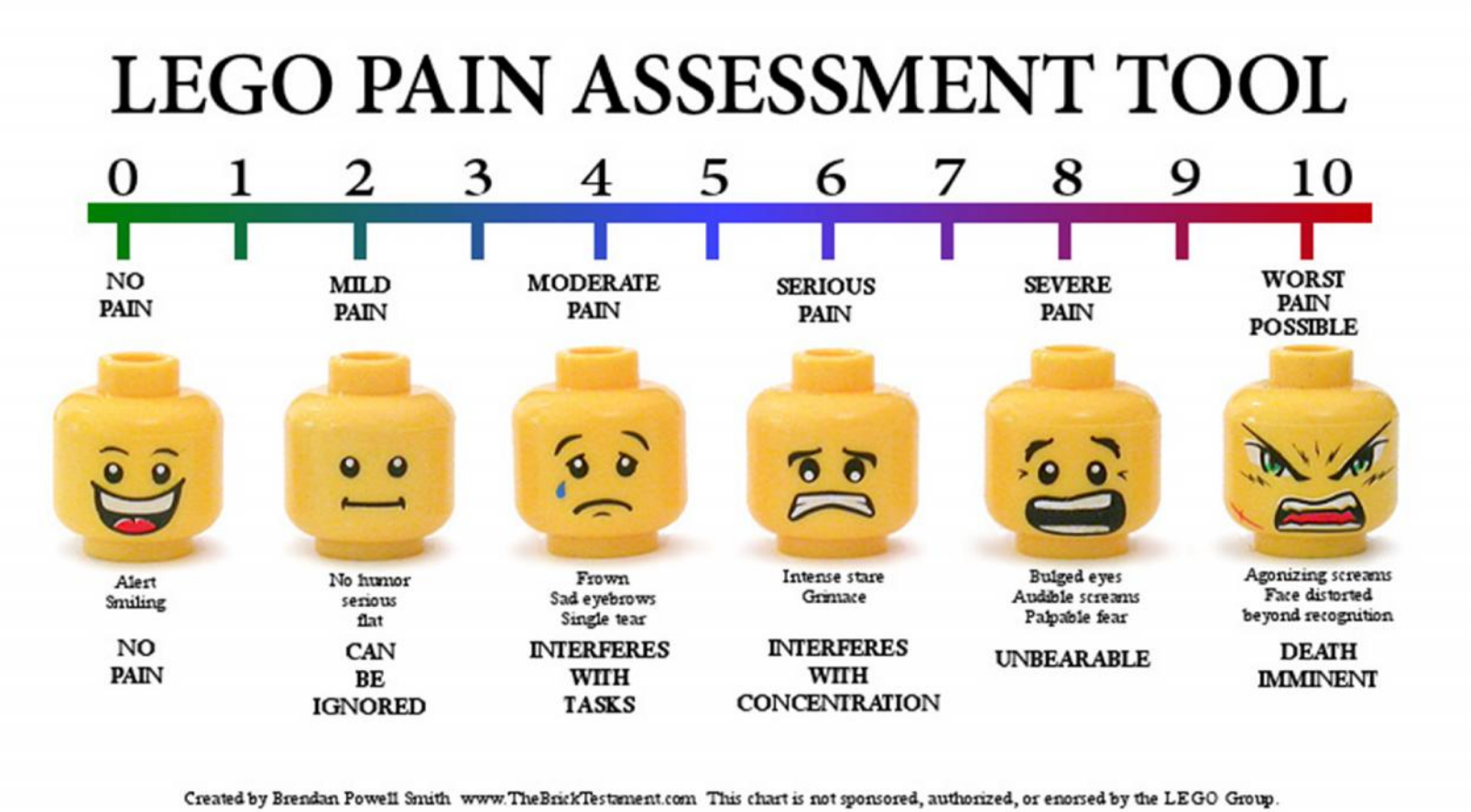
50 patients got a new pain-killer; 50 similar patients got an older drug. Mean self-reported pain scores were 4.1 for the new drug and 4.5 for the old drug.

Example: drug testing

Statistical inference addresses:

- How can we estimate the difference in the effect of the drugs?
- How can we quantify the precision of that estimate?
- Is it possible that the apparent difference is due to chance?

But how bad is it?



Statistical inference

Effect size: usually a single number, ideally comparable across studies.

Confidence interval and/or standard error: quantifies the precision of the estimate.

p-value: indicates whether the apparent effect might be due to chance.

Statistical inference

Effect size: by far the most important!

Confidence interval and/or standard error:
a distant second.

p-value: an even more distant third.

Part of the problem

Too many papers/research report p-values as if they were the most important part.

And bury the effect size!

Girls Good at Math, But Teachers May Make Them Anxious About it

Jan. 26, 2010

By LAUREN COX
ABC News Medical Unit

Elementary school teachers who are nervous about doing math in public may be unintentionally grooming young girls to think they are bad at math -- even if they are just as capable as the boys in their class -- a small study suggests.

Psychologists tested 17 female **elementary school teachers** on a widely-used Mathematics Anxiety Rating Scale to see how comfortable the teachers felt doing math in public situations. The researchers also measured the ability -- **and self confidence** -- of 52 boys and 65 girls starting in the first or second grade.

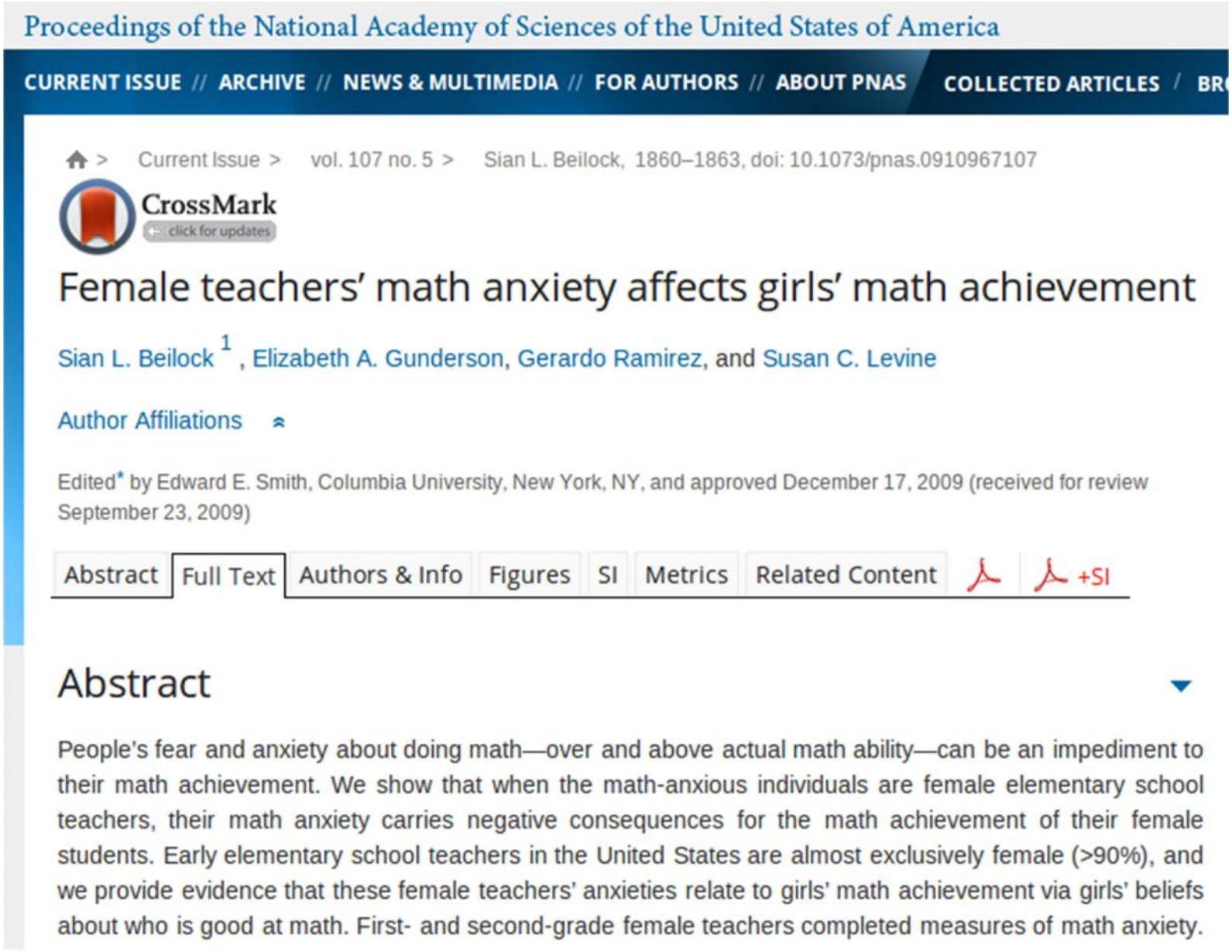
In the first few months of the school year, there was no measurable **difference between boys' and girls' math capabilities**. But by the end of the year, girls who had math-anxious teachers were more likely to do worse on math achievement tests than boys in their class -- and worse than girls who were in classrooms with teachers who felt confident in math.

"The more anxious a teacher is in the situation, the

This sounds bad

...girls who had math-anxious teachers were more likely to do worse on math achievement tests than boys in their class -- and worse than girls who were in classrooms with teachers who felt confident in math.

"The more anxious a teacher is in the situation, the more likely girls are going to pick up on this," said [the lead author]



http://www.pnas.org/content/107/5/1860.abstract

You have to work for it

Regression analysis established that teachers' math anxiety had a significant negative effect on girls' math achievement at the end of the school year ($\beta = -0.21$, $t = -2.17$, $P = 0.034$).

Teacher math anxiety scores ranged from 1.6 to 4.2 of a possible 5.

Beginning of the year: girls 69, boys 29
End of year: girls 72, boys 21

And the answer is...

The difference in outcome for the most anxious teacher, compared to the least anxious, is 0.55 points on a test where the mean is ~100 points.

And it’s possible that this apparent effect is due to chance.

Don’t make me work!

Report effect size prominently.
In terms I can interpret.
And provide context for comparison.

http://lifechangingx.com/wp-content/uploads/2014/07/no-working.png

Let’s get to it

Launch Jupyter.

Load effect_size.ipynb

Read, do the first exercise.

Stop when you get to STOP HERE.

What have we learned?

Obvious measure of effect size is difference in means, in centimeters.

Relative difference, as a percentage, might be useful, but you might have to choose the denominator.

Back to it

Read, do the second exercise.

Read about Cohen’s d.

Play with the interactive widget.

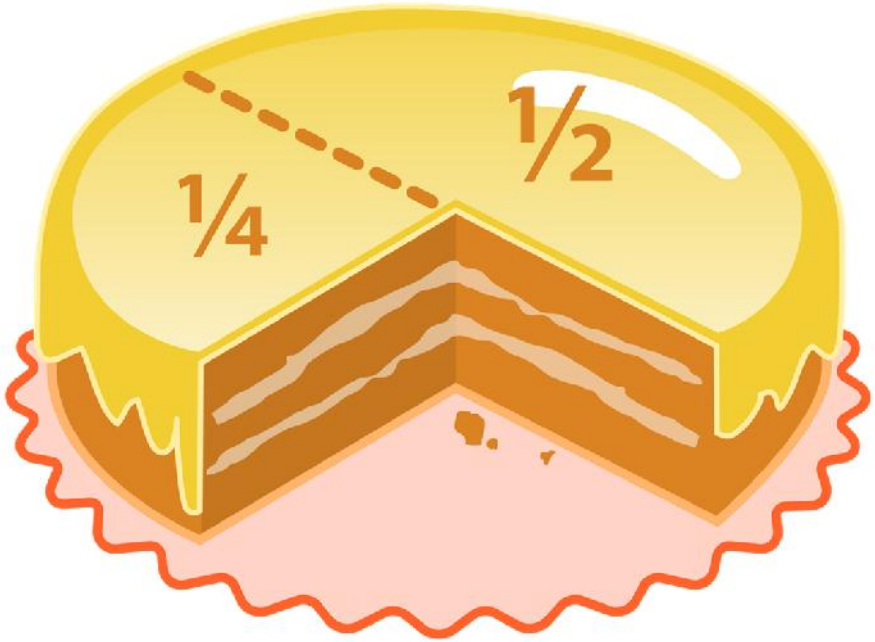
What have we learned?

Alternatives for comparing distributions: overlap, misclassification, probability of superiority.

Cohen’s effect size: symmetric, standardized, comparable across studies.

Effect size #2

Differences in proportions.



http://upload.wikimedia.org/wikipedia/commons/thumb/4/4f/Cake_fractions.svg/2000px-Cake_fractions.svg.png

Leap

Learning Early About Peanut Allergy

LEAP is no longer accepting study participants

LEAP Study Results

For Media

About ITN

Other Allergy Studies

Resources

Understanding Clinical Trials

Patient Rights & Informed Consent

Children & Clinical Studies

The results of the Immune Tolerance Network's (ITN) "Learning Early About Peanut" (LEAP), discussed on February 23, 2015 at the American Academy of Allergy, Asthma & Immunology Annual Meeting and published in the *New England Journal of Medicine*, demonstrate that consumption of a peanut-containing snack by infants who are at high-risk for developing peanut allergy prevents the subsequent development of allergy. The LEAP study, designed and conducted by the ITN with additional support from FARE and led by Professor Gideon Lack at Kings College London, is the first randomized trial to prevent food allergy in a large cohort of high-risk infants.

Peanut allergy is an aberrant response by the body's immune system to harmless peanut proteins in the diet. The prevalence of peanut allergy has doubled over the past 10 years in the US and other countries that advocate avoidance of peanuts during pregnancy, lactation, and infancy. The LEAP study was based on a hypothesis that regular eating of peanut-containing products, when started during infancy, will elicit a protective immune response instead of an allergic immune reaction.

Changes in proportions

"Of the children who avoided peanut, 17% developed peanut allergy by the age of 5 years. Remarkably, only 3% of the children who were randomized to eating the peanut snack developed allergy by age 5. Therefore, in high-risk infants, sustained consumption of peanut beginning in the first 11 months of life was highly effective in preventing the development of peanut allergy."

http://www.leapstudy.com/leap-study-results

What’s the headline number?

"All the News That's Fit to Print."

The New York Times.

THE WEATHER

NEW YORK, MONDAY, NOVEMBER 11, 1918. TWENTY-FOUR PAGES. TWO CENTS

ARMISTICE SIGNED, END OF THE WAR! BERLIN SEIZED BY REVOLUTIONISTS; NEW CHANCELLOR BEGS FOR ORDER; OUSTED KAISER FLEES TO HOLLAND

SON FLEES WITH EX-KAISER

Hindenburg Also Believed to be Among Those in His Party.

ALL ARE HEAVILY ARMED

ON THEIR WAY TO DE STEEG

Kaiser Fought Hindenburg's Call for Abdication; Failed to Get Army's Support in Keeping Throne

By GEORGE RENWICK.

Kaiser Shivered as He Signed Abdication

BERLIN TROOPS JOIN REVOLT

Reds Shell Building in Which Officers Vainly Resist.

THRONGS DEMAND REPUBLIC

GENERAL STRIKE IS BEGUN

Socialist Chancellor Appeals to All Germans To Help Him Save Fatherland from Anarchy

WAR ENDS AT 6 O'CLOCK THIS MORNING

The State Department in Washington Made the Announcement at 2:45 o'Clock.

ARMISTICE WAS SIGNED IN FRANCE AT MIDNIGHT

Terms Include Withdrawal from Alsace-Lorraine, Disarming and Demobilization of Army and Navy, and Occupation of Strategic Naval and Military Points.

WASHINGTON, Monday, Nov. 11, 2:48 A. M.—The armistice between Germany, on the one hand, and the allied Governments and the United States, on the other, has been signed.

The State Department announced at 2:45 o'clock this morning that Germany had signed.

The department's announcement simply said: "The armistice has been signed."

The world war will end this morning at 6

GERMAN DYNASTIES BEING WIPED OUT

MORE WARSHIPS JOIN THE REDS

Difference in percentage

“Eating peanuts decreases chance of allergy by 14 percentage points.”

OR

“Avoiding peanuts increases chance of allergy by 14 percentage points.”

Difference in percentage

Percentage points are not created equal:

From 3% to 17% is a big deal.

From 43% to 57% might not be.

From 0.1% to 14.1% is huge!

https://www.mint.com/blog/wp-content/uploads/2013/04/savings-percentage.jpg

Percentage change

“Administering peanuts decreases allergy rates by 83%.”

That’s good, but...

Percentage change

“Administering peanuts decreases allergy rates by 83%.”

That’s good, but...

“Avoiding peanuts increases allergy rates by 467%.”

Percentage change

Solved the first problem, but:

- Effect size depends on how we define treatment and control.
- And people get confused by percent changes in percentages.

Odds ratio

$odds = p / (1 - p)$

$p_1 = 0.03$
$p_2 = 0.17$

$o_1 = 0.0309$
$o_2 = 0.2048$

$OR = o_1 / o_2 = 0.151$

Odds ratio

“Eating peanuts decreases allergy rates (OR=0.15).”

OR

“Avoiding peanuts increases allergy rates (OR=6.6).”

Odds ratio

OR is most common in practice.

But not symmetric.

And tricky to plot and do arithmetic with.

Log odds ratio

“Eating peanuts decreases allergy rates (LOR=-0.82).”

OR

“Avoiding peanuts increases allergy rates (LOR=0.82).”

Log odds ratio

In many ways, LOR is the best way to work with probabilities and changes in probability.

And it’s comparable between studies.

The obvious downside is lack of familiarity.

Axolotl

http://divaboo.info/

Summary

Treatment	Difference in rate	Percent change	Odds ratio	Log odds ratio
Administer peanuts	-14 points	-83%	0.15	-0.82
Withhold peanuts	+14 points	+467%	6.6	+0.82

Summary

Rule #1: Choose a measure of effect size that is meaningful in context.

Rule #2: The estimated effect size is the most important result. Everything else is auxiliary.

Standardized measures are comparable across studies (but sometimes violate Rule #1).

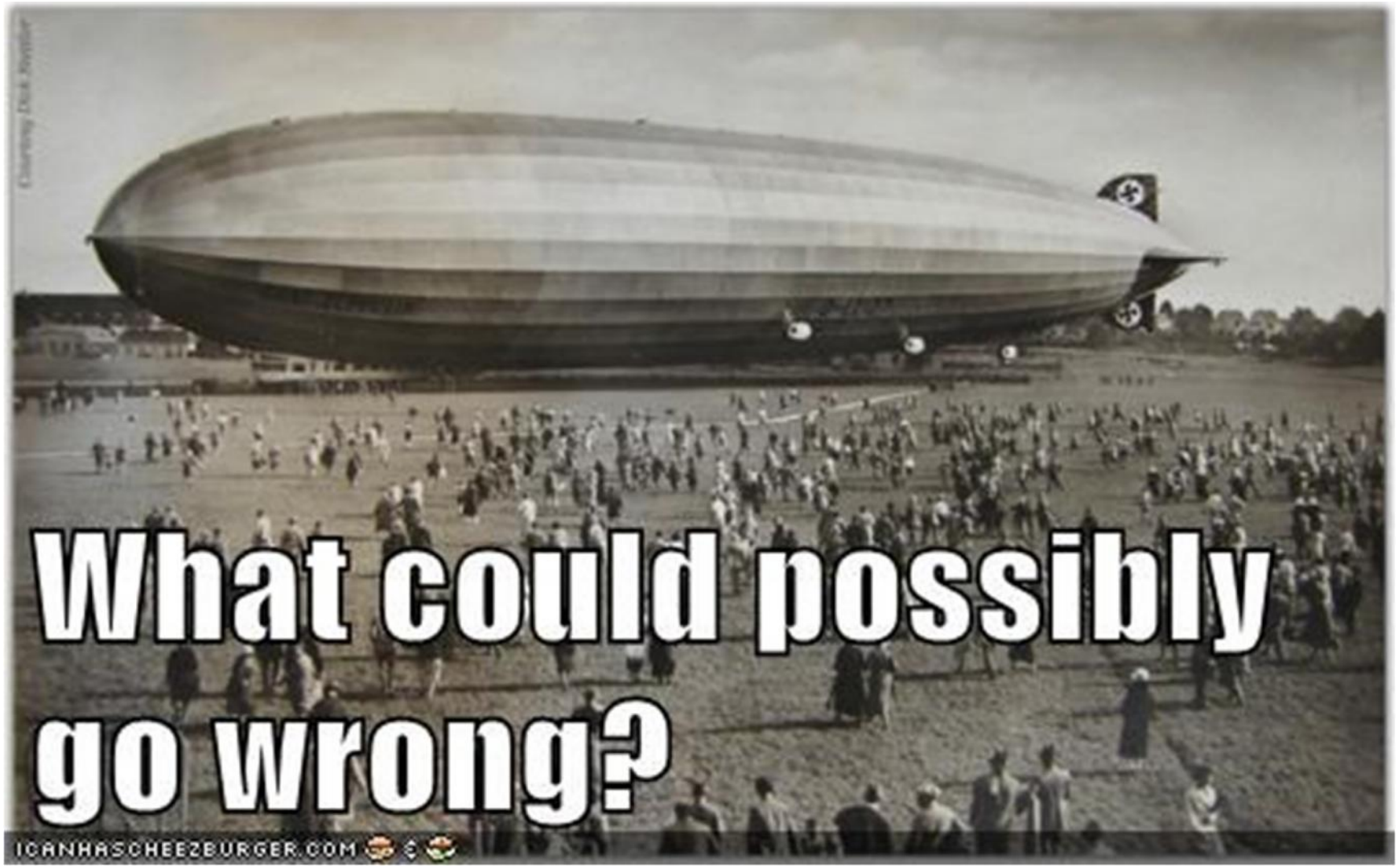
http://images.forwallpaper.com/files/images/2/271a/271a3e17/598269/coffee-break-yet.jpg

Part Two: Quantifying precision

Suppose you want to know the distribution of weight for adults in the India.

You do a telephone survey using random numbers from the phone book.

Whoever answers the phone, you ask for their weight and write down the answer.


What could possibly go wrong?
ICANHASCHEEZBURGER.COM

Sampling bias

Ideally the sample should be representative.

Or if some groups are oversampled, there should be no relationship between group membership and what you measure.

Sampling bias

Poor people might have more adults per phone.

Rich people might have unlisted numbers.

Unemployed people might be more likely to be home when you call.

And weight is probably related to wealth and employment.

Measurement error

1. Clothes and shoes.
2. Uncalibrated scale.
3. Misremembered.
4. Misreported.
5. Misrecorded.
6. Miscoded (kg or lb? 999=NA?).

Random error

Suppose you only call 3 people. By chance, you might get 3 heavy people.

Even for large samples, each possible sample yields different measurements.

Three kinds of error

Sampling bias: hard to quantify.

Measurement error: sometimes quantifiable.

Random error: relatively easy to quantify.

It’s common to discuss the first two qualitatively, and quantify the third.

Edward Tufte
@EdwardTufte

Follow

Random error is but one of 20 threats to learning from data.
Disproportionate attention because it is only threat math-modeled.
#datascience

RETWEETS	LIKES
32	10

2:49 PM - 3 Nov 2013

Back to Jupyter

Load up `sampling.ipynb`

Read, run the examples, interact with the widgets, do the exercises.

Stop when you get to STOP HERE.

What have we learned?

One way to quantify variability is to run lots of simulated experiments and compute sample statistics.

The distribution of sampling statistics is the sampling distribution, which we can use to compute SE and CI.

What have we learned?

As the sample size increases, SE and the width of the CI get smaller.

We can use this framework to compute sampling distributions for other statistics: coefficient of variation, 90th percentile, etc.

However...

We cheated!

In the examples, we knew the actual population distribution.

How do we “simulate lots of experiments” without knowing the population distribution?

Modeling and simulation

Use the sample to make a model of the population.

Use the model to simulate more experiments.

Resampling

One way to model the population and simulate experiments:

1. Treat the sample as a population.
2. Draw new samples with the same size (with replacement).

Resampling

One way to model the population and simulate experiments:

1. Treat the sample as a population.
2. Draw new samples with the same size (with replacement).

Back to Jupyter

Back to `sampling.ipynb`

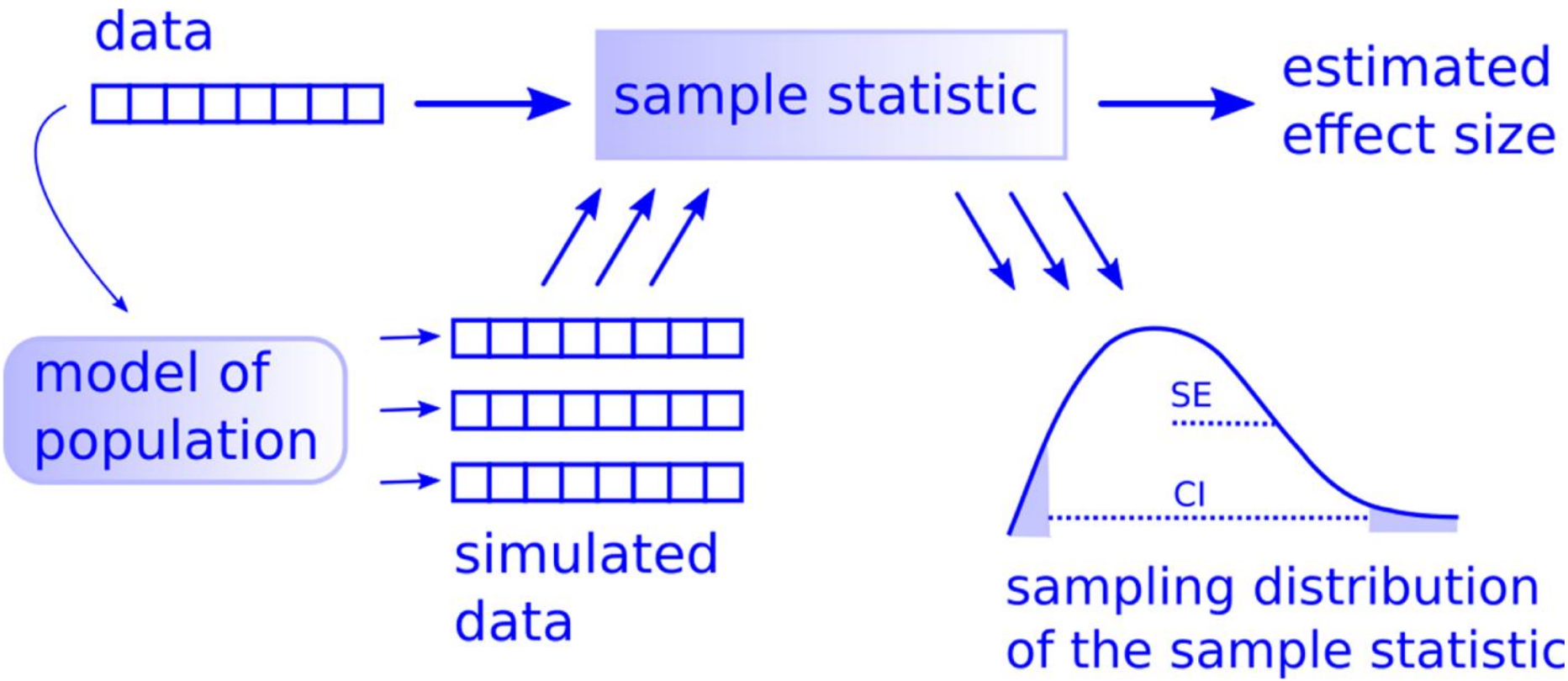
Read, run the examples, interact with the widgets, do the exercises.

Stop when you get to STOP HERE.

Summary

1)Use the sample to model the population.
2)Use the model to generate new samples.
3)Compute the sampling distribution of whatever statistic you want.
4)Report SE or CI, or both (but be clear about which it is).

Resampling


data
sample statistic
estimated
effect size
model of
population
simulated
data
SE
CI
sampling distribution
of the sample statistic

Template pattern

For people who like design patterns.

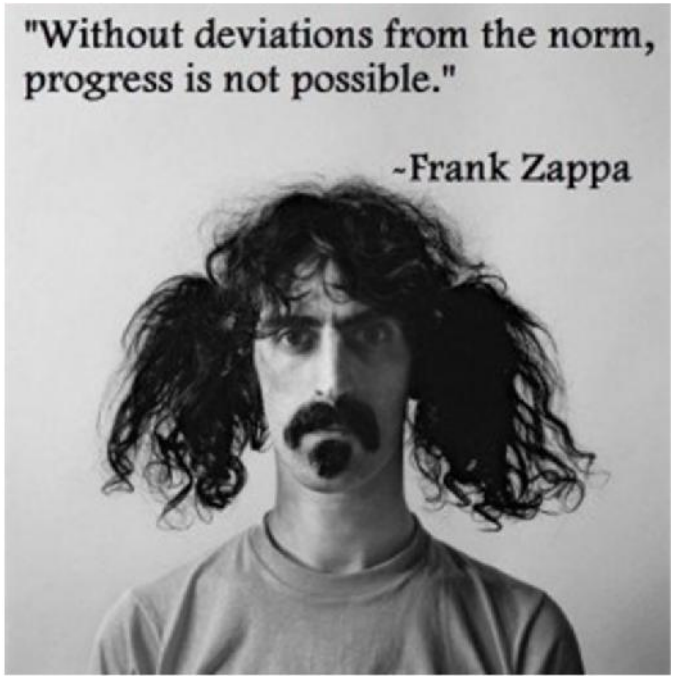
If we have time

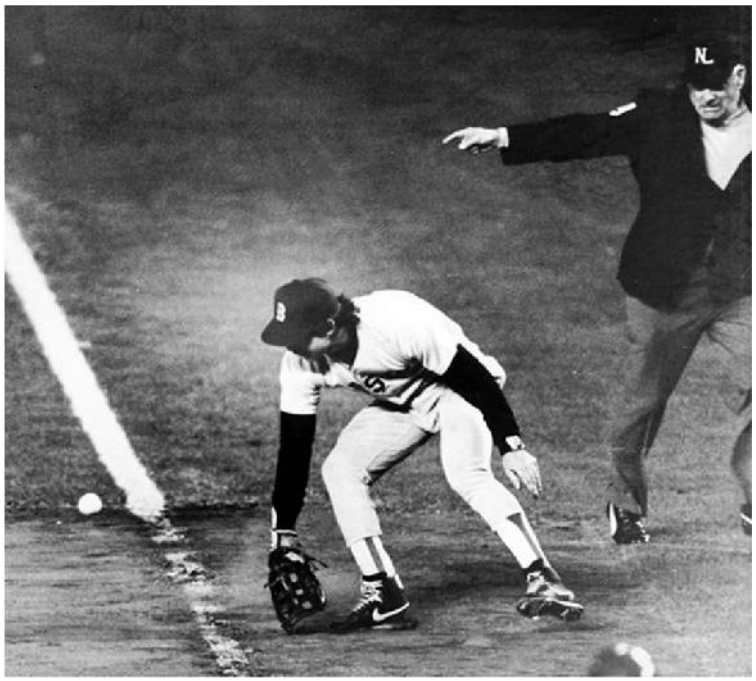
Back to `sampling.ipynb`

Read and run Part Three.

Review question

What's the difference between standard deviation and standard error?





Review question

Standard deviation: Summary statistic that describes a population. The SD of adult male height is 7.7 cm.

Standard error: Quantifies the precision of an estimate. The mean adult height in the BRFSS sample is 178.5024 cm. The SE of this estimate is 0.00034 cm.

Review question

In the BRFSS sample, the average male height is 178.5024 cm, and the 95% CI is

[178.5018, 178.5031] cm

What is the probability that the actual population average is in this range?

Review question

In the BRFSS sample, the average male height is 178.5024 cm, and the 95% CI is

[178.5018, 178.5031] cm

What is the probability that the actual population average is in this range?

We don’t know.

Interpreting CIs

If there is no sampling bias, or measurement error, or modeling error, there is a 95% chance that the 95% CI you computed contains the actual population statistic.

There are two objections to this statement...

The black hole

"The actual population mean is not a random variable, so you are not allowed to talk about the probability that it falls in an interval. It either does or it doesn't."

http://en.wikipedia.org/wiki/Black_hole#/media/File:PIA16695-BlackHole-Corona-20130227.jpg

The relevant objection

Sampling bias, measurement error, and modeling errors are inevitable.

The CI only accounts for random error. As sample size increases, the CI gets smaller, and other errors dominate.

Therefore, CIs from large samples are LESS likely to contain the actual value.

Summary

SE and CI quantify variability due to random sampling.

Resampling is a simple and general way to compute them.

Don’t forget about other sources of error.

Part Three: Hypothesis testing

I left it for last because it is the least important.

If it was wiped from the face of the earth tomorrow, we would be better off.

But you might be curious, or you might be required to do it.

So...

NHST

Null Hypothesis Significance Testing.

The logic is similar to proof by contradiction.

http://weheartit.com/entry/group/18241158

Proof by contradiction

I’m trying to prove A.

Assume temporarily that A is false.

That assumption leads to something impossible.

Therefore A must be true.

NHST

I observe an apparent effect in a sample.

Assume temporarily that there is no effect in the population.

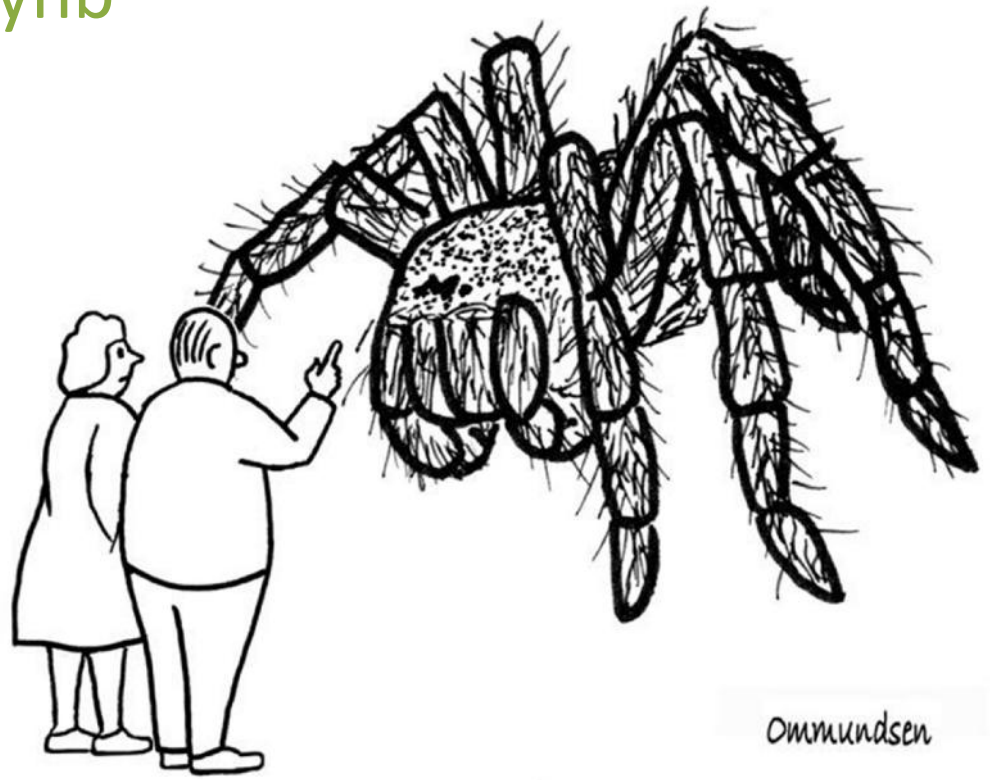
Compute the probability of seeing the effect in a sample if there is no effect in the population.

If it’s small, the effect is probably not random.

To the notebook!

Load up hypothesis.ipynb

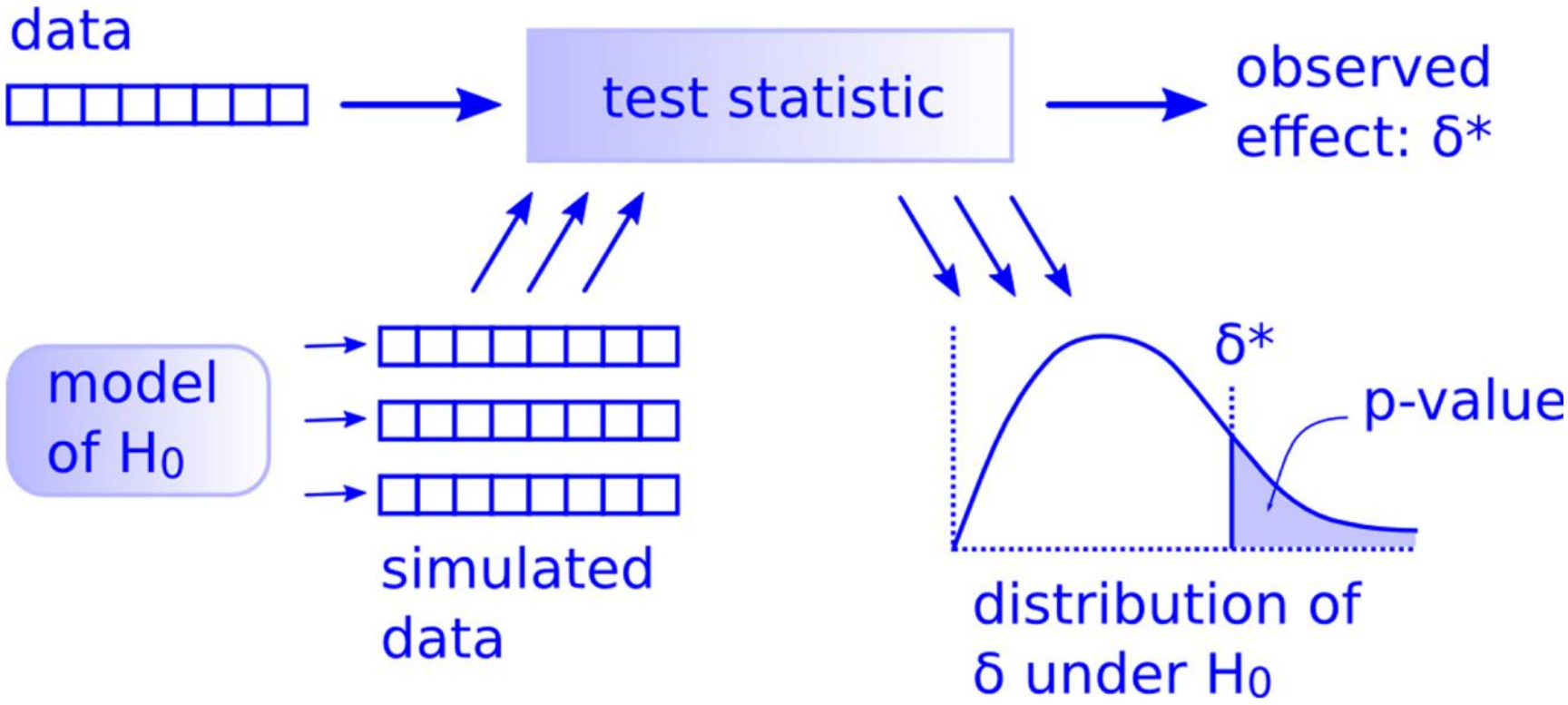
Read and run
until you get
to STOP HERE.



"I've narrowed it to two hypotheses:
it grew or we shrunk."

http://eugenieteasley.com/wp-content/uploads/2014/11/Hypothesis-pic.jpg

There is only one test



What have we learned

Test statistic: whatever number you choose to quantify the magnitude of the effect.

Null hypothesis: a model of a hypothetical world where the apparent effect is not real.

https://xkcd.com/892/

Permutation

If the null hypothesis is that two groups are the same, we can simulate it by pooling the groups and shuffling.

That’s called a permutation test.

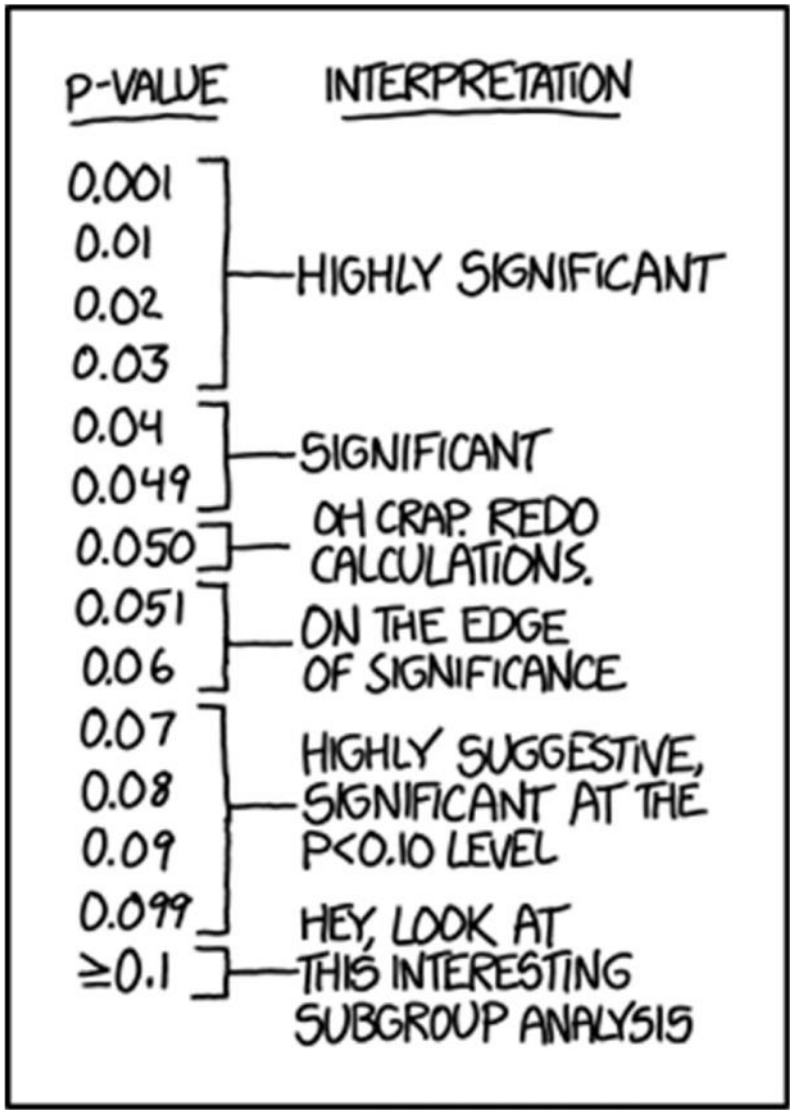
p-value

The p-value is the probability that the test statistic, under the null hypothesis, exceeds the observed value.

If it’s small, you can conclude that the apparent effect is probably not due to chance*.

* Some sticklers object to this phrasing, but I stand by it.

Interpreting p-values the xkcd way



https://xkcd.com/1478/

More seriously

Interpret the order of magnitude of the p-value:

More than 10%, plausibly due to chance.

Less than 1%, probably not.

In between, borderline.

HypothesisTest

The nice thing about the computational approach is that it handles other models and other test statistics.

The HypothesisTest class represents this general framework.

Back to hypothesis.ipynb

Reminder

NHST can rule out one explanation, random sampling, but not:

- Sampling bias,
- Measurement error,
- Confounding variables,
- Fraud,
- Honest mistakes,
- etc.

Summary

Effect size is important.

SE and CI quantify error due to randomness (but not other sources of error).

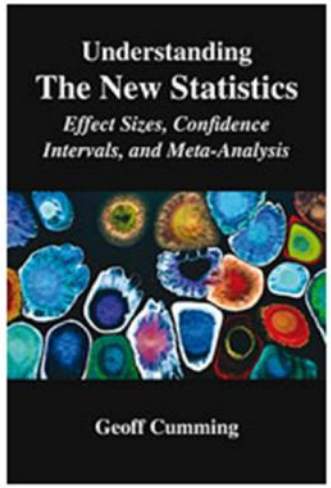
p-values indicate whether an effect might be due to chance (but that’s often not the thing we should worry about).

The New Statistics: Estimation for better research

The book

Cumming, G. (2012). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge

- Explains estimation, with many examples.
- Designed for any discipline that uses statistical significance testing.
- For advanced undergraduate and graduate students, and researchers.
- Comes with free ESCI software.
- May be the first evidence-based statistics textbook.
- Assumes only prior completion of any intro statistics course.
- See the *dance of the confidence intervals*, and many other intriguing things.



http://www.psychologicalscience.org/index.php/members/new-statistics