Lecture 10 – November 27, 2017

{ topics for today }

The replication crisis

P-Hacking → responsible for replication crisis

{ last week }

- Roles that values play in definitions of psychiatric disorders
- Implications for people diagnosed with such disorders

{ exam format }

12 short answer questions (short paragraphs)

Lectures and readings → topics on class website

The Replication Crisis

- Going to look at issues to do with the use of statistics
- New context for values here – critical role in sci investigation → it has just begun to be explored
- How does big data change science → big data revolution
- This course was not about what is moral behaviour is but about histomology…
    - What we know and how we came to know about it
- Science stands in better epistemic foundation than other kinds of knowledge about the world
- But status of knowledge generated by sci inquiry is too simple
- Value-free ideal → scientists works independent from social and political pressures
- This is too simple… not borne out in the real world
- Value-free ideal fails to obtain in numerous informal ways
- How does the influence of values on scientists doing science effect the content/status of knowledge generated by the scientific project?
- Values have some role to play in science
- Objectivity (truth-tracking nature) → is there such thing that can be maintained in science?
- Society ↔ Science
    - Values pointed in both directions
- We thought about climate change, funding, psychiatry and now statistics
- All related in the sense that they are concerned with the role of uncertainty in decision making
- Climate change: there is decent amount of scientific information but it comes with uncertainties; we don't know what actions to take
    - There are false positives and negatives
    - Entrance point for values
- Funding: how to divide funds, central difficulty is uncertainty related to what a given kind of research project is going to produce (any large outcomes that may affect people's lives?)
    - Hard to see how, before you do research, if it is going to affect society
    - Entrance point for values
- Psychiatry: there is some uncertainty in this domain as well (less clear); the way symptoms are realized, aka biological mechanisms, are not really understood; plays a significant role in how to think about whether or not one has a mental disorder
- Pattern → all have an entrance point of values closely related to instances where science cannot give us certain knowledge…
- However, science is usually the best way of trying to figuring out what to do in the face of uncertainty
- But values still play a role

Slide 1

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets PPV = $(1 - \beta)R/(R$

- Active scientists have gone ahead and said there is a real crisis that we are facing
- This crisis was realized recently
- The crisis is our ability to replicate scientific results
- Look at the title → "why most published research findings are false" (2005)
- Startling claim in a scientific research
- That means it was refereed by professional scientists and judged it to be well argued and supported
- What is causing this circumstance?
- Is it a true claim?
- Most sci research claims to establish false conclusions
- Why should we worry?

Slide 2

THE REPLICATION CRISIS: IT HAS RECENTLY BEEN DISCOVERED THAT A SURPRISINGLY SMALL AMOUNT OF REPORTED SCIENTIFIC RESULTS ARE REPRODUCIBLE.

- People do experiments in lab, they follow the sci methods and standard statistics
- They find the statistically-significant conclusion
- They write a paper
- They send it to journal
- Journal sees data, results, conclusion
- Paper gets published
- Expanding jobs
- If the same people in lab try to reproduce experiment using same methodology…
- And they reproduce the same calculations…
- What they find is that the effect they initially reported goes away
- This is very bad; why is it going on?

Slide 3

WHY IS REPRODUCIBILITY IMPORTANT?

- We want this bc it's an indication that we have a truly predicted theory
- Not just having spurious correlations in data and reporting those rather than real data in the world
- Reproducibility is an important metric/standard
- If something is reproducible then it is at least possible we latched on to some real feature of the world
- We at least know that we are getting something about the world right
- If one day, the theory has a consequence and the other day it doesn't, then it doesn't suggest what is really going to happen in the world

Slide 4

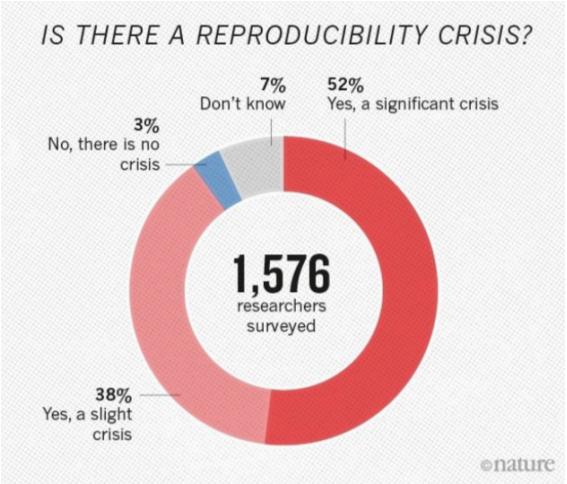
REPRODUCIBILITY IS AN IMPORTANT INDICATION THAT SCIENTIFIC RESULTS ARE GETTING THE WORLD RIGHT

Slide 5

If an effect is real we should be able to reproduce it whenever the conditions relevant for the effect obtain.

- If there is a real feature of the world we pick up on when we do science, we should be able to reproduce it
- PV = MRT
- It would be surprising if you weren't able to reproduce that pattern between different variables in lab
- It has been reproduced so many times in lab!
- So if fails, it probably had to do with incapability of a researcher
- Replication crisis → people are reporting facts as facts like PV = MRT
    - Reproduce the result by reproducing the data, but the effect is not the same

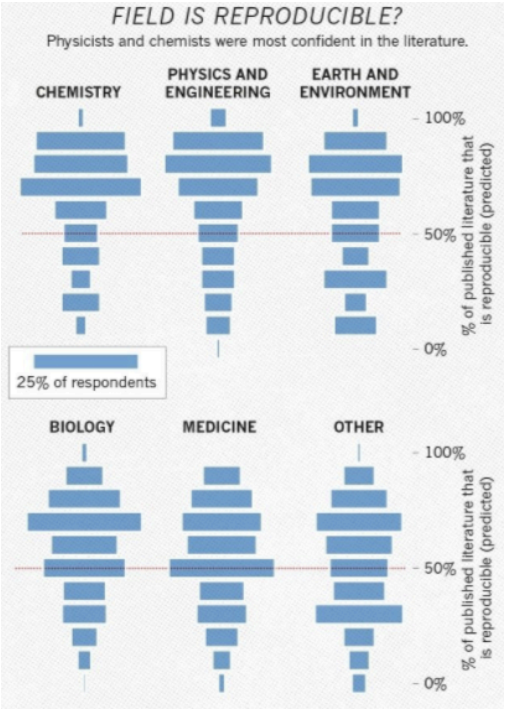
Slide 6 + 7



- Nature survey (important sci journal)
- Unreproducible research is public
- Assignment 15 scientists and asked about attitude about reproducibility

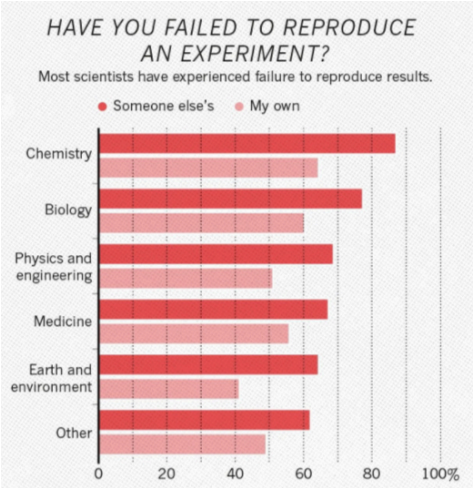
- Q1: Is there a reproducibility crisis?
- 3% say no
- 7% do not know
- 38% yes, slight crisis
- 52% yes, significant crisis
- There is a problem that stands out in different areas of sci inquiry

Slide 8



- Asked about their discipline
- Percentage of published literature that is reproducible?
- Chemistry →
    - Most popular claims is 70% is reproducible – some people think more
    - Little people claim less is reproducible
- Medicine
    - Most popular value: 50%

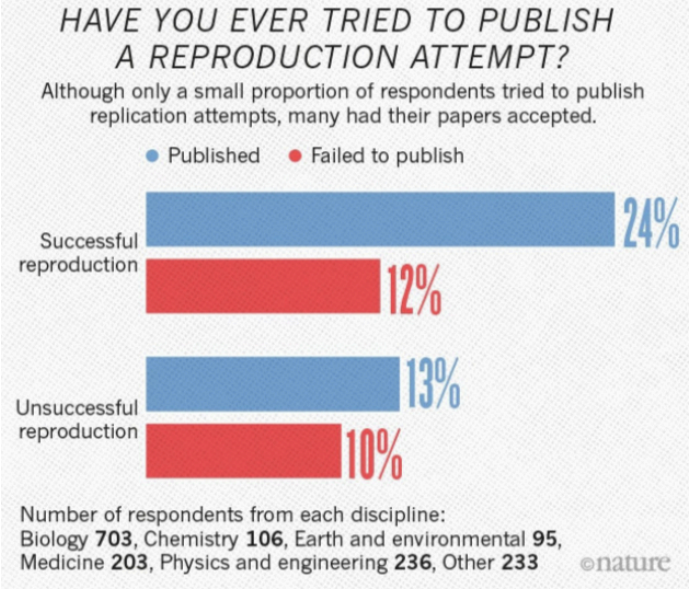
Slide 9



- Asked about own personal experience attempting to replicate their results

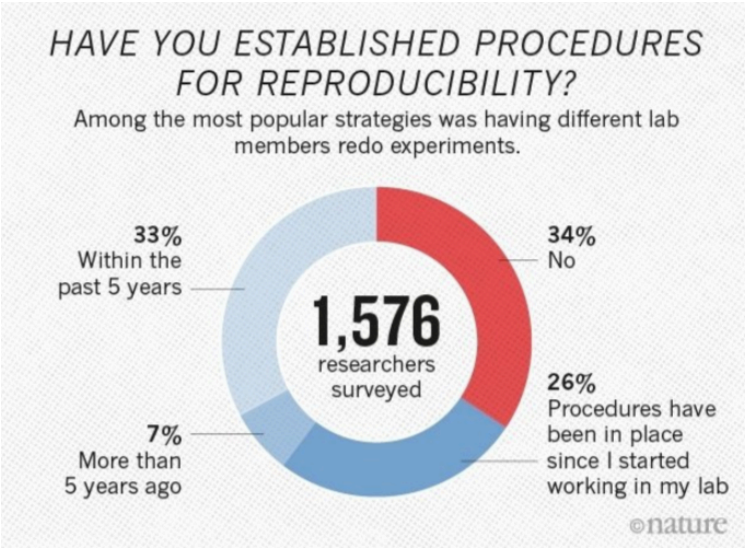
- Look at subheading and labels
- Chemistry → 90% of respondents failed to reproduce someone else's results, 65% not being able to reproduce results for their own publications
- Biology → somewhat better than chemistry
- Not about perceived problem about reproducibility, but researchers have attempted to reproduce results and failed to do so

Slide 10



- Have you ever tried to publish a reproduction attempt?
- Successful reproduction → 24% published, 12% failed to publish
    - They did not send to journal or they did but it was rejected
- Unsuccessful reproduction → 13% published, 10% failed to publish
- What do we do about this?

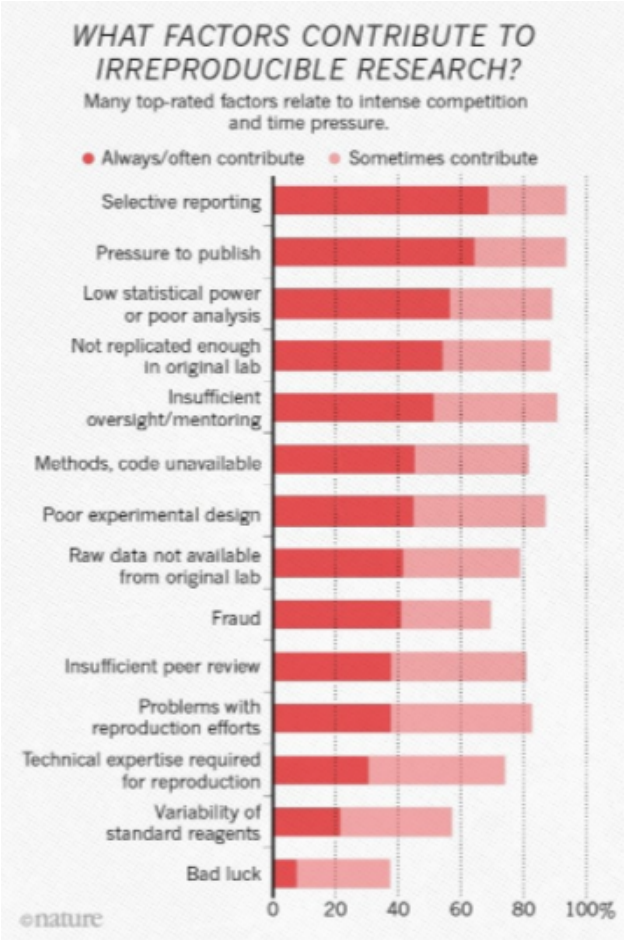
Slide 11



- Have you established procedures for reproducibility?
- Whether or not in their lab, have they adopted a set of procedure to make research more likely to be reproducible
- 1/3 say no → no dedicated procedure to enhance reproducibility of their research

- 1/3 say yes within last 5 years → after realization of crisis
- 1/4 say yes
- Most popular strategy was having diff lab members do the experiments

Slide 12



- What factors contribute to irreproducible research?
- Most often was selective reporting (sometimes or always contributes) → what is this?
    - Varies from context
    - P-hacking is an instance of it
        - Searching for different statistical correlations in the data available
        - Don't publish instances of failed correlations that are not stat significant and only those that are stat significance
- Pressure of publish → important
- Low stat power or poor analysis, fraud

Slide 13

- Selective reporting and low stat power are supposedly related to one another

Slide 14

WHAT DOES ANY OF THIS HAVE TO DO WITH VALUES?

- Where are the values in this story about current state of research?
- Pressure to public → not much to do with actual science but something influencing the scientists
- Values are coming in because external pressures – external to scientific project of figuring out what is really going on in the world, imposed on those doing the research

Slide 15

PRESSURE TO PUBLISH

Slide 16

WHERE DOES THIS COME FROM

- You need them in your resume in order to survive as a scientist

Slide 17

SCIENTISTS ARE PEOPLE TOO

- Supposed to conform to a specific set of norms
- But scientists are people
- They are subject to all kinds of pressures including all other people in the world

Slide 18

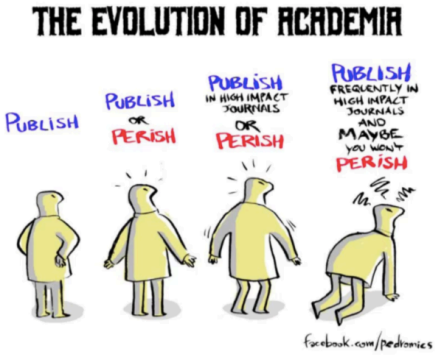
THEY RESPOND TO PRESSURES NOT ENTIRELY INTERNAL TO THE SCIENTIFIC PROCESS

Slide 19

THEY COMPETE FOR POSTDOCS, PROFESSORSHIPS, GRANT MONEY, PRESTIGE

- Competition never ends
- Selects for people willing to participate

Slide 20



Slide 21

THE PERSONAL DESIRES OF INDIVIDUAL SCIENTISTS AFFECT REPRODUCILILITY

- This is how values come in
- Their desire to succeed is affecting reproducibility

Slide 22

IS THIS A NEW CONNECTION BETWEEN SCIENCE AND VALUES?

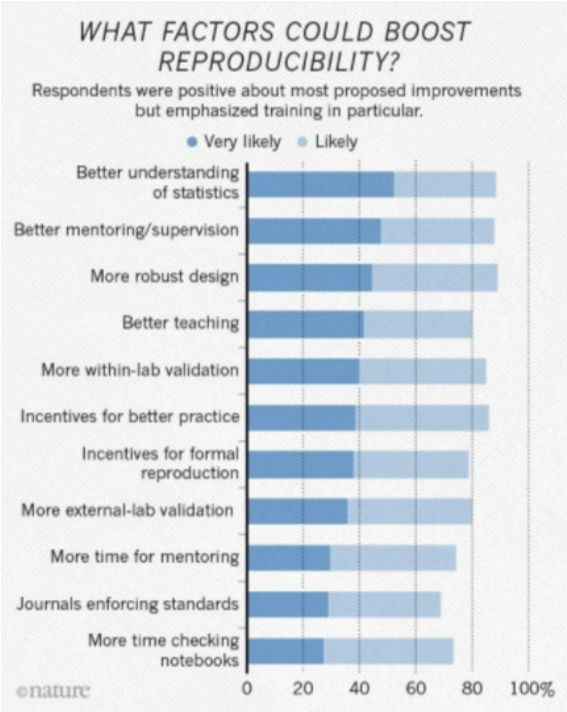
- Or is it related to other connections between science and values we've talked about?

Slide 23

THE PRESSURE GETS MANIFESTED IN THE WAY THAT STATISTICS GET USED/MISUSED

- Not just that people are publishing or fraudulently reporting results
- Something more subtle
- Has to do with the way people misuse stats

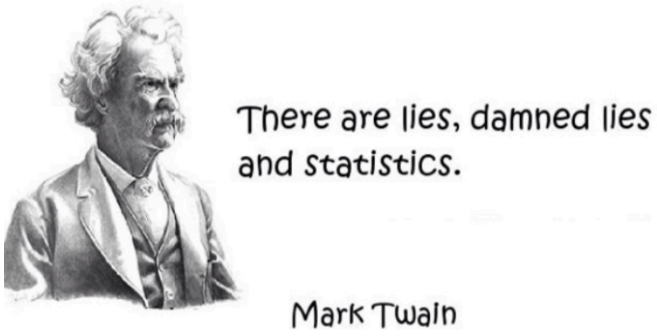
Slide 24



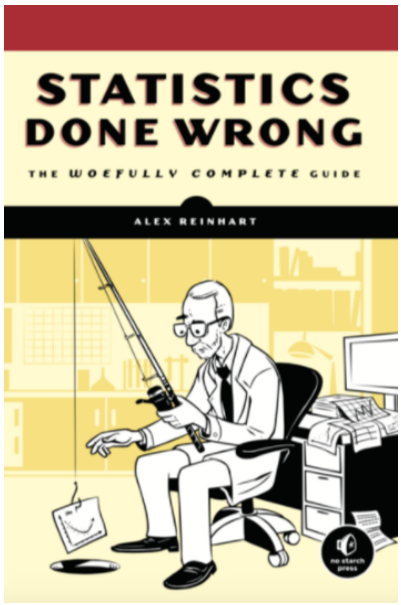
- What factors could boost reproducibility?
- The first thing reported is not reduced fraud
- It is that people need a better understanding of stats
- That would probs improve reprod
- It is the researchers themselves not understanding how to do statistical tests properly when determining which results to publish

P-Hacking

Slide 1



Slide 2



- Material that's follows comes from this book

Slide 3

- Not p-Hack data but emphasises the data
- Look at correlation
- Surprising result
- Two things that do not have anything to do together
- But looks like it is correlated to one another
- Example of spurious kinds of correlations that are published in reproduced results

Slide 4

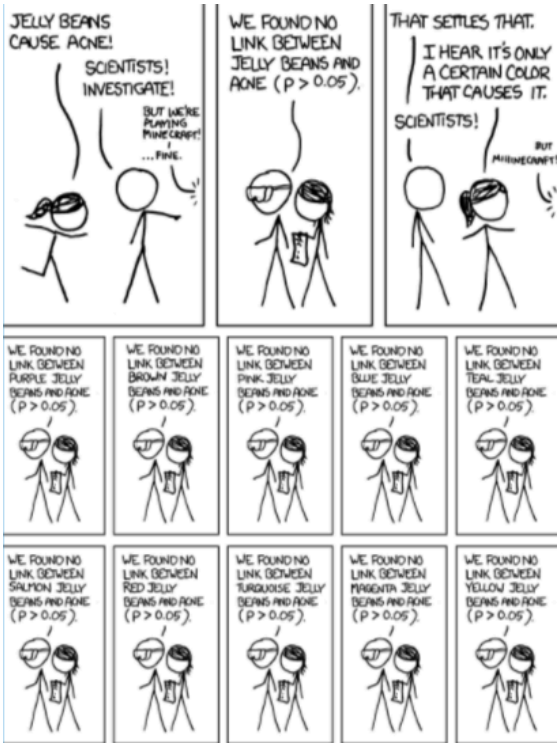
P-HACKING REFERS TO A PARTICULAR KIND OF INAPPROPRIATE MANIPULATION OF STATISTICAL DATA

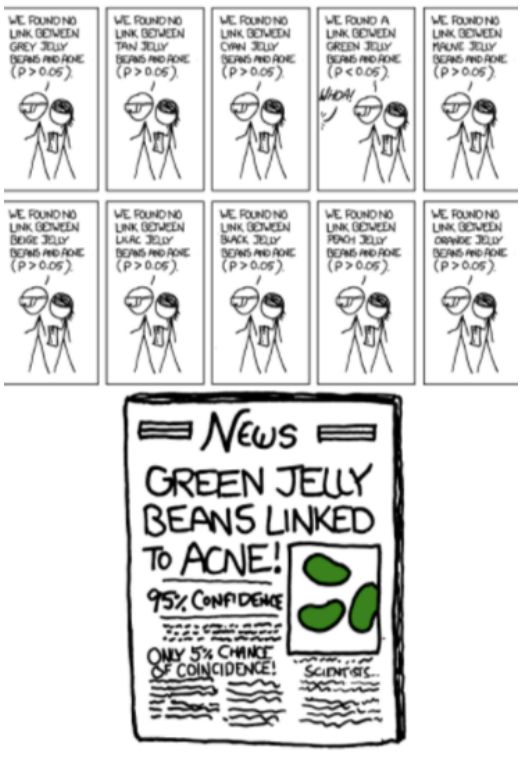
Slide 5

The basic idea: observe many different hypothesized statistical correlations. By chance, some will be surprising. Report those but not all of the other unsurprising results.

- Just run a whole bunch of test for stat significant
- By chance some of them will be surprising
- The report those surprising instances but not the other ones
- What happened in the spider and spelling graph?
- Somebody probably took a distribution of a whole bunch of variables
- Just tested them all against each other
- For a bunch of them, found no correlation bc just random variables (have nothing to do with each other)
- Only surprising, if you don't see all the other non-surprising ones

Slide 6 + 7





- Two scientists worrying about jellybeans causing acne
- Found a p-value greater than 0.05, so no justify that there is no link
- Somebody says it's the colour of the jellybeans
- No stat significance for all the colours except green ($p<0.05$)
- Joke: you are looking at a bunch of stat tests
- Eventually (just by chance) (just by nature of what p-values means), you should expect to get spurious correlations some percentage of the time (5% of them time) (95% confidence)
- If you do the 100 tests, 5 of them will be just a coincidence

Slide 8

p-Values: a measure of how surprising a result is. They are frequently used to determine whether or not an effect is statistically significant.

- P-values are sometimes misinterpreted
- This is a standard applied when people make decisions about publishing certain results

Slide 9

Reinhart: Does one medecine work better than another? We use statistics to make judgments about these kinds of differences. We will always observe some difference due to luck and random variation, so statisticians talk about statistically significant differences when the difference is larger than could easily be produced by luck. So first we must learn how to make that decision.

- About stats going wrong
- In sci asking questions about differences between different variables
- Do certain chemicals cause cancer?
- We uses stats to make judgments about these differences
- We will always observe some difference due to luck and random variations
- Just the Feature of world
- Stat sig diff → when the difference is larger and could easily be produced by luck?
- How do we know?
- What is the diff between having good reason to things that some correlations are so surprising (we think we are getting evidence for real effect in the world) as opposed to having reason to just that we are observing something resulting from chance or random variation?

Slide 10

Suppose you're testing cold medicines. Your new medicine promises to cut the duration of cold symptoms by a day. To prove this, you find 20 patients with colds, give half of them your new medicine, and give the other half a placebo. Then you track the length of their colds and find out what the average cold length was with and without the medicine.

- Testing cold medicine → new meds cut down the duration of sickness by a day?
- 20 patients; half of them get new med and have get placebo
- Track the length of their cold
- Same statistical methodology applied to testing Higgs boson and across many other sciences
- To know if something is stat significant

Slide 11

But not all colds are identical. Maybe the average cold lasts a week, but some last only a few days. Others might drag on for two weeks or more. It's possible that the group of 10 patients who got the genuine medicine in your study all came down with really short colds.

How can you prove that your medicine works, rather than just proving that some patients got lucky?

- Not all colds are the same → colds can last 2 days, a week, 3 weeks, etc.
- There could be a random chance that all of the people that had the meds had a short cold in the first place and all the others assigned with the placebo got a long cold

Slide 12

Statistical hypothesis testing provides the answer. If you know the distribution of typical cold cases—roughly how many patients get short colds, long colds, and average-length colds—you can tell how likely it is that a random sample of patients will all have longer or shorter colds than average.

By performing a hypothesis test (also known as a significance test), you can answer this question: "Even if my medication were completely ineffective, what are the chances my experiment would have produced the observed outcome?"

- Stat hypothesis testing
- If you know the normal distribution of typical cold cases in the absence of med intervention
- You can tell how likely a random set of patients will have longer/shorter colds than average
- By performing the test, you can answer a question → if the med was doing nothing, what is the likelihood that the experiment yielded the results it did

Slide 13

If you test your medication on only one person, it's not too surprising if her cold ends up being a little shorter than usual. Most colds aren't perfectly average. But if you test the medication on 10 million patients, it's pretty unlikely that all those patients will just happen to get shorter colds. More likely, your medication actually works.

- If randomly pick one person, it has gotta be one of the numbers in the distribution
- It may be more likely they may get a cold somewhere in the middle
- But you may get one that doesn't have the average cold which is not surprising
- Most colds aren't the average value
- Test the med on 10M patients (large sample size), it is unlikely that all patients will have just shorter cold
- Or that the med is actually doing something
- That is why you need a large sample size to make sure results are not due to random chances

Slide 14

Scientists quantify this intuition with a concept called the p value. The p value is the probability, under the assumption that there is no true effect or no true difference, of collecting data that shows a difference equal to or more extreme than what you actually observed.

- There is a numerical way of quantifying basic picture of what is going on
- The p-value → assume that the med has no true effect
- Give the cold med to sample population, you see some shortening of the cold

- The p-value gives us the probability of having that shortening of the cold duration (for the population getting med) for a more significant reduction than the one we actually observed

Slide 15

So if you give your medication to 100 patients and find that their colds were a day shorter on average, then the p value of this result is the chance that if your medication didn't actually do anything, their average cold would be a day shorter than the control group's by luck alone.

As you might guess, the p value depends on the size of the effect—colds that are shorter by four days are less common than colds that are shorter by just one day—as well as on the number of patients you test the medication on.

- P-value depends on the size of effect → read passage
- Also depends on the number of patients you test them on

Slide 16

Remember, a p value is not a measure of how right you are or how important a difference is. Instead, think of it as a measure of surprise. If you assume your medication is ineffective and there is no reason other than luck for the two groups to differ, then the smaller the p value, the more surprising and lucky your results are—or your assumption is wrong, and the medication truly works.

- P-value does not measure how right you are
- It is a measure of how surprise you should be by the results you got
- How unusual is it with respect to normal expectations (of what the normal distribution of the variable should be)
- Assume that if ineffective
- The smaller the pi-value, the more surprising or lucky your results are
- OR your assumption is wrong and your med truly works
- Basically people try to show that It is *so* unlikely that the result in question of change just by chance that it should be justified in thinking that the difference we are seeing is due to the medication is working

Slide 17

How do you translate a p value into an answer to this question: "Is there really a difference between these groups?" A common rule of thumb is to say that any difference where $p < 0.05$ is statistically significant. The choice of 0.05 isn't because of any special logical or statistical reasons, but it has become scientific convention through decades of common use.

- Look at question
- $P < 0.05$ → statistically significant

- P-value is less than a particular number
- 0.05 is not scientific or special → just a convention (maybe need revision due to problems of reproducibility)
- Determining the p-value depends on the number of subjects we test
- To get lower p-values (situations where it is more unlikely that there results were due to luck) requires more data to consider
- But very expensive to have huge sample size
- Higgs Bosons → looking for small deviations
- Need a huge number of events to look at in order to produce large sample sizes
- Took years of collecting data before being able to conclude that there were sufficient stats
- To know if there was a stat sig diff b/w the data and what one would expect if there was no Higgs boson
- But that is expensive
- You would need people there to collect data for years
- Just like how you can't get too many animals for testing (limited)
- Under pressure to publish and you have low sample sizes → you might find other ways where you can produce negative results but some of which are positive results (is not speaking good methodology)
- Pressure form publication and costs of performing becomes realized

Slide 18

p is a measure of surprise, with a smaller value suggesting that you should be more surprised. It's not a measure of the size of the effect. You can get a tiny p value by measuring a huge effect— "This medicine makes people live four times longer"—or by measuring a tiny effect with great certainty.

Slide 19

Statistical significance does not mean your result has any practical significance. As for statistical insignificance, it doesn't tell you much. A statistically insignificant difference could be nothing but noise, or it could represent a real effect that can be pinned down only with more data.

- What does Reinhart mean by that?
- If you don't get sat sig result → doesn't tell you anything at all whether there is an effect involved
- You only get info about what you should believe, if you get a stat sig result

Slide 20

There's no mathematical tool to tell you whether your hypothesis is true or false; you can see only whether it's consistent with the data. If the data is sparse or unclear, your conclusions will be uncertain.

- Uncertainty coming back and messing with the way that values enter in

Slide 21

A p value of 0.05 corresponds to a 5% chance of mistakenly rejecting the null hypothesis (i.e. treating the effect as not due to chance even though it is).

So if you do enough hypothesis tests you will end up with some results in this 5%.

- Gold standard of sat sig is 0.05
- 5% chance of mistakenly rejecting the null hypothesis
- You will end up with some results from that 5% (not a real effect in the world but due to chance)
- P-hacking comes to be a part of the story
- Do a whole bunch of tests and report results in this 5% and makes it look like there is an effect
- Just reporting spurious correlations (no real significance)

Slide 22

Is everything we eat associated with cancer? A systematic cookbook review[1–3]

Jonathan D Schoenfeld and John PA Ioannidis

ABSTRACT

Background: Nutritional epidemiology is a highly prolific field. Debates on associations of nutrients with disease risk are common in the literature and attract attention in public media.

Objective: We aimed to examine the conclusions, statistical significance, and reproducibility in the literature on associations between specific foods and cancer risk.

Design: We selected 50 common ingredients from random recipes in a cookbook. PubMed queries identified recent studies that evaluated the relation of each ingredient to cancer risk. Information regarding author conclusions and relevant effect estimates were extracted. When >10 articles were found, we focused on the 10 most recent articles.

Results: Forty ingredients (80%) had articles reporting on their cancer risk. Of 264 single-study assessments, 191 (72%) concluded that the tested food was associated with an increased ($n = 103$) or a decreased ($n = 88$) risk; 75% of the risk estimates had weak ($0.05 > P \geq 0.001$) or no statistical ($P > 0.05$) significance. Statistically significant results were more likely than nonsignificant findings to be published in the study abstract than in only the full text ($P < 0.0001$). Meta-analyses ($n = 36$) presented more conservative results; only 13 (26%) reported an increased ($n = 4$) or a decreased ($n = 9$) risk (6 had more than weak statistical support). The median RRs (IQRs) for studies that concluded an increased or a decreased risk were 2.20 (1.60, 3.44) and 0.52 (0.39, 0.66), respectively. The RRs from the meta-analyses were on average null (median: 0.96; IQR: 0.85, 1.10).

Conclusions: Associations with cancer risk or benefits have been claimed for most food ingredients. Many single studies highlight implausibly large effects, even though evidence is weak. Effect sizes shrink in meta-analyses. *Am J Clin Nutr* 2013;97:127–34.

and such discrepancies in the evidence have fueled hot debates (9–12) rife with emotional and sensational rhetoric that can subject the general public to increased anxiety and contradictory advice (13, 14). One wonders whether this highly charged atmosphere and intensive testing of food-related associations may create a plethora of false-positive findings (15) and questionable research practices, especially when the research is highly exploratory, the analyses and protocols are not preregistered, and the findings are selectively reported. It was previously shown in a variety of other fields that "negative" results are either less likely to be published (16–21) or misleadingly interpreted (19, 22). Studies may spuriously highlight results that barely achieve statistical significance (15, 23) or report effect estimates that either are overblown (24, 25) or cannot be replicated in other studies (24, 26, 27).

To better evaluate the extent to which these factors may affect studies investigating dietary risk factors for malignancy, we surveyed recently published studies and meta-analyses that addressed the potential association between a large random sample of food ingredients and cancer risk of any type of malignancy.

SUBJECTS AND METHODS

Random ingredient selection

We selected ingredients from random recipes included in *The Boston Cooking-School Cook Book* (28), available online at http://archive.org/details/bostoncookingsch00farmrich. A copy of the book was obtained in portable document format and viewed by using *Skim* version 1.3.17 (http://skim-app.sourceforge.net). The recipes (*see* Supplementary Table 1 under "Supplemental data" in the online issue) were selected at random by generating random numbers corresponding to cookbook

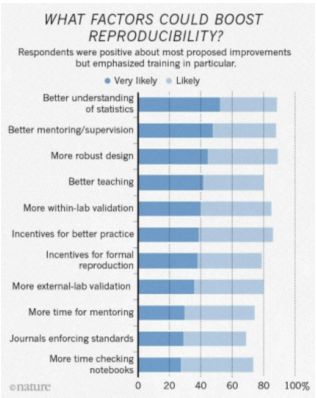
- One week vitamin D is good for you but next week, you might that its bad for you liver
- Contradictory messages what is good or not for our health → nutritional epidemiology
- Someone picked up on this theme and came up with the title
- Probably due to stat manipulations
- Should this guide my actions in the world in what to eat or not eat
- Randomized: looked at cookbook and picked up 50 ingredients

- People tried to identify p-hacking (not going out and checking if ingredients cancer) but look at all the papers that have been published that claim stat significant data about particular ingredients causing cancer → check whether or not we should believe those results (is it due to chance or real effect)
- This generates a peak curve; shows the distributions of published p-values
- You can come up with the argument that majority of published research (question of whether nutrient causes cancer) is p-hacked →not to generate fraudulent conclusion
- Just to get a p-value less than 0.05, so that they can publish it and move on in their career
- Majority is spurious correlation
- That does not mean that there aren't correlation between nutrients and cancer
- It just means that research done on the topic and published used it as a means for us to believe that these things caused cancer → are not reliable and statistical meaningful results

Slide 23

- Prof did a little p-hacking here
- What is being tested is whether or not the republicans and democrats has an influence on economic performance → probably a claim that will get lodged in a debate
- If you could show there was an improvement if a republication or democrat in office…
- You need to be really careful about statistics; extremely easy to p-hack
- Moral of story: p-hack data is extremely misleading

Slide 24



- What factors could boost reproducibility?
- To better the understanding of stats (most critically important to reduce it)
- Once you go out into world, in lab there are diff people in lab
- Underappreciated → being a person that is good at stats does not make you the most popular
- That attitude will get relaxed and change to some extent
- Learn stats no matter what you end up doing

CASE FOUR → @1:40:00