Critiques of NHST

Previously

- NHST

- Type I and Type II errors

- Power

1. Define H_0 and H_1.

2. Choose your α level.

3. Collect data.

4. Define your sampling distribution using your null hypothesis and either the knowns about the population or estimates of the population from your sample.

5. Calculate the probability of your data or more extreme under the null. (To get the probability, you'll need to calculate some kind of standardized score, like a z-statistic.)

6. Compare your probability (p-value) to your α level and decide whether your data are "statistically significant" (reject the null) or not (fail to reject the null).

Where does that leave us?

- What kind of science has NHST produced?

- The replication crisis

- What the future holds

Critiques of NHST

Many will describe the current "replication crisis" or "reproducibility crisis" or "open science movement" as tracing its beginnings to 2011, mostly due to:

- Bem. (2011). Feeling the future.
- Simmons, Nelson, & Simonsohn. (2011). False-positive psychology.

But the reality is that NHST has always had its critics. And it's not for lack of eloquence that they have been ignored...

"The textbooks are wrong. The teaching is wrong. The seminar you just attended is wrong. The most prestigious journal in your scientific field is wrong." – Ziliak and McCloskey (2008)

"... surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students" – Rozeboom (1997)

"Statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution" – Schmidt and Hunter (1997)

"What's wrong with [NHST]? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" – Cohen (1994)

"... an instance of a kind of essential mindlessness in the conduct of research" – Bakan (1966)

"... despite the awesome pre-eminence this method has attained in our journals and textbooks of applied statistics, it is based upon a fundamental misunderstanding of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research" – Rozeboom (1960)

What kind of mess have we gotten ourselves into?

- $p < .05$ as a condition for publication

- Publication as a condition for tenure

- Novelty as a condition for publication in top-tier journals

- Institutionalization of NHST

- High public interest in psychological research

- Unavoidable role of human motives: fame, recognition, ego

What kind of science have we produced?

- $p < .05$ as a primary goal

- Publication bias: "Successes" are published, "failures" end up in file drawers

- Overestimation of effect size in published work

- Underestimation of complexity (why did the failures occur?)

- Underestimation of power

- Inability to replicate

- Settling for vague alternative hypotheses: "We expect a difference"

What kind of science have we produced?

- Dichotomous thinking (based on p): research either "succeeds" or "fails" to find the expected difference

- No motivation to pursue failures to reject the null

- Harvesting (mostly) the low-hanging fruit in science to publish quickly and often.

- Weak theory

  - Low precision ("differences" are enough)
  - No non-nil null hypotheses
  - Weak, slow progress as a science
  - see paper/tweet by Eiko Fried

Focusing on p -values

Imagine rolling a die.

- What's the probability you roll a 2?

    - $P(2) = 1/6 = 16.7\%$

- If you roll the die twice, what's the probability that you get a 2 at least once? 30.6%

- If you roll the die 5 times, what's the probability that you get a 2 at least once? 59.8%

Roll the die enough times, and you'll get a 2 eventually. Significance testing when the null is true is like rolling a 20-sided die.

False Positive Psychology

Simmons et al. (2011) pointed out that each study is not a single roll of the die.

Instead, each study, even those with a single statistical test, might represent many rolls of the die.

- **Researcher degrees of freedom:** Decisions that a researcher makes that change the statistical test.
  - Examples:
    - Additional dependent variabiles
    - Tests with and without covariates
    - Data peeking (testing effect as data comes in and stopping when result is significant)

Each time we see how a decision affected our result, we are rolling the dice again.

Table 1. Likelihood of Obtaining a False-Positive Result

Researcher degrees of freedom	Significance level		
	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Questionable Research Practices (QRPs)

- *p*-**hacking**
- HARKing
- cherry picking
- fishing
- data dredging

Collecting data or analyzing your data in different ways until non-significant results become significant.

Prior to 2011, this was common practice. In fact, it was often taught as best practices.

- "Explore your data."
- "Understand your data."
- "Test sensitivity..."

What's wrong with this?

Questionable Research Practices (QRPs)

- p-hacking
- **HARKing**
- cherry picking
- fishing
- data dredging

*H*ypothesizing *A*fter the *R*esults are *K*nown

You analyze data and find a significant result (might be unexpected), and post-hoc come up with a hypothesis. Importantly, you then *report* the findings as though this has been the case all along.

What's wrong with this?

Questionable Research Practices (QRPs)

- p-hacking
- HARKing
- **cherry picking**
- fishing
- data dredging

Select/report only data/findings that support your hypothesis. If your data/findings do not support it, you hide it away in the "file drawer"

What's wrong with this?

Questionable Research Practices (QRPs)

- p-hacking
- HARKing
- cherry picking
- **fishing**
- data dredging

Look at a ton of different combinations of variables to find something significant.

What's wrong with this?

Questionable Research Practices (QRPs)

- p-hacking
- HARKing
- cherry picking
- fishing
- **data dredging**

Takes fishing to an extreme, but usually more synonymous with "data mining". Usually no hypothesis in mind. Just go crazy in some database or "big data" set, looking for things that pop up as significant.

What's wrong with this?

Where we're at

The publication of False Positive Psychology, following the claim by Ioannidis (2005) that as many as half of published findings are false prompted researchers to take a second look at the "knowns" in our literatures.

If we can demonstrate these "known" effects, then we're ok. Our effects are most likely true.

And if that had happened, we probably wouldn't have two lectures in this class dedicated to problems with NHST and how to address them.

Just repeat your experiment...

The *inability* to replicate published research has been viewed as especially troubling.

- This has been a long-standing concern, but the poster child is undoubtedly "Estimating the reproducibility of psychological science" by the Open Science Collaboration (Science, 2015, 349, 943).

Only 36% of the studies were replicated, despite high power and claimed fidelity of the methods.

	Replications $P < 0.05$ in original direction	Percent	Effect size comparison					
			Mean (SD) original effect size	Median original df/N	Mean (SD) replication effect size	Median replication df/N	Average replication power
Overall	35/97	36	0.403 (0.188)	54	0.197 (0.257)	68	0.92
JPSP, social	7/31	23	0.29 (0.10)	73	0.07 (0.11)	120	0.91
JEP:LMC, cognitive	13/27	48	0.47 (0.18)	36.5	0.27 (0.24)	43	0.93
PSCI, social	7/24	29	0.39 (0.20)	76	0.21 (0.30)	122	0.92
PSCI, cognitive	8/15	53	0.53 (0.2)	23	0.29 (0.35)	21	0.94

Have you ever tried to replicate a study?

Why is it so hard to replicate?

- Poor understanding of context necessary to produce most effects

  - We do not recognize the boundary conditions of effects especially when the limiting conditions are kept constant

- Incomplete communication of the necessary conditions

  - Akin to reading just the first few ingredients for a recipe and then trying to duplicate the dish.

Why is it so hard to replicate?

Sparse communication fosters belief by others that effects are simpler and easier to produce than they really are.

The reality is that key elements have been left out:

- specific methodological or analytic details

- and the tests run before and after the ones that were published.

Calls for change have included:

- Effect size estimates and CI but not p

- Require exact replication attempts as a condition of publication and funding.

- Require pre-registration.

- Publish everything and let meta-analysis sort it out.

What kind of challenges will need to be addressed?

- Institutionalization of CI and ES but not p

- Journals must change their values

- How do we shift media values? Replication is not "sexy."

- How do we shift academic values? What will be the impact on tenure? How do we re-calibrate "productivity?" Citation indices?

Is Psych bullshit?

The inability to replicate much work has been viewed as a singular failure of psychology as a science.

- But is that the best way to view it?

As data, the replication crisis, it is potentially quite informative and suggests some importance features of psychology as a science.

At a minimum, it tells us we don't understand a phenomenon as well as we think we do. That is hardly a failure of science, unless we don't take the next steps to resolve our ignorance.

Can science progress?

Failures to replicate are surprising, even troubling. We think of them as unfortunate occurrences. Bad luck even. But should they be viewed that way?

Surprises in science have a long history of playing a key role in scientific progress.

What can (or should) failures to replicate tell us?

A good laboratory, like a good bank or corporation or government, has to run like a computer. Almost everything is done flawlessly, by the book, and all the numbers add up to the predicted sums. The days go by. And then, if it is a lucky day, and a lucky laboratory, somebody makes a mistake . . . something is obviously screwed up, and then the action can begin. The next step is the crucial one. If the investigator can bring himself to say, "But even so, look at that!" then the new finding, whatever it is, is ready for the snatching. What is needed for progress to be made, is the move based on the error . . . The capacity to leap across mountains of information to land lightly on the wrong side represents the highest of human endowments. --Lewis Thomas (1974) *The Medusa and the Snail*

A failure to replicate should be viewed as particularly interesting.

> It is time to insist that science does not progress by carefully designed steps called "experiments" each of which has a well-defined beginning and end. Science is a continuous and often disorderly and accidental process. A first principle not formally recognized by scientific methodologists: when you run onto something interesting, drop everything else and study it. --B. F. Skinner (1968) *A case history in scientific method*

The ability not only to look straight at what you want to see, but also to watch continually, through the corner of your eye, for the unexpected. I believe this to be one of the greatest gifts a scientist can have. Usually, we concentrate so much upon what we intend to examine that other things cannot reach our consciousness, even if they are far more important. This is particularly true of things so different from the commonplace that they seem improbable. Yet, only the improbable is really worthy of our attention! If the unexpected is nevertheless found to be true, the observation usually represents a great step forward. --Hans Selye (1964) *From Dream to Discovery*

"The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!', but 'That's funny...'" -Isaac Asimov

[Science] needs a better word for "error" . . . Or maybe "error" will do after all, when you remember that it came from an old root meaning to wander about, looking for something. -Lewis Thomas

Wrapping Up

- You can't solve all jobs with a wrench

- NHST is just one of the tools in your toolbelt.

- Effect sizes, confidence intervals, replications are all also tools to help give us a clearer picture.

Next time...

Open Science (and methods for improving your inference)

If time, R stuff!