Back to Basics

And the mess we've created

Null Hypothesis Significance Testing

(NHST)

- A hypothesis is a statement about a population
- Null Hypothesis (H_0) is the hypothesis against which the research result is tested
  - It is often 0 (nil hypothesis) but doesn't have to be
  - In words, "there is no relationship between X & Y" or "the intervention has no effect on the outcome" or "there is no difference between two values"
  - In math:

$$H_0 : cor_{XY} = 0$$

$$H_0 : x_1 - x_2 = 0$$

$$H_0 : x_1 - x_2 = 4$$

Null Hypothesis Significance Testing

(NHST)

- We work with probabilities. In order to do so, you need to cover all possible events
- The alternative hypothesis (H_1 or H_A) is every possible event not represented by the null hypothesis

$$H_0 : \mu = 4$$

$$H_1 : \mu \neq 4$$

$$H_0 : \mu < -4$$

$$H_1 : \mu \geq -4$$

Null Hypothesis Significance Testing

Steps

1. Define H0 and H1 (null and alternate hypotheses)

2. Choose your alpha level (what are you willing to call “unlikely”?)
   - Less than 5%? — Less than 1% — Less than 10%?
   - This acts as your threshold for making decisions

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 alpha level and decide whether your data are "statistically significant" (reject the null) or not (fail to reject the null).

Example

Step 1: Define H_0 and H_1

- H_0: Ratings of women come from the same population as the ratings of men; means are the same
- H_1: Ratings of women do not come from the same population as ratings of men; means are different

$$H_0 : \mu_{men} = \mu_{women}$$

$$H_1 : \mu_{men} \neq \mu_{women}$$

Example

Step 2: Set α

Define what “unlikely” means

- Is a 5% chance or less considered unlikely?
- Is a 10% chance or less considered unlikely?

Most often, $\alpha = .05$

Our hypothesis is 2-sided; that is, we have not specified a *direction*

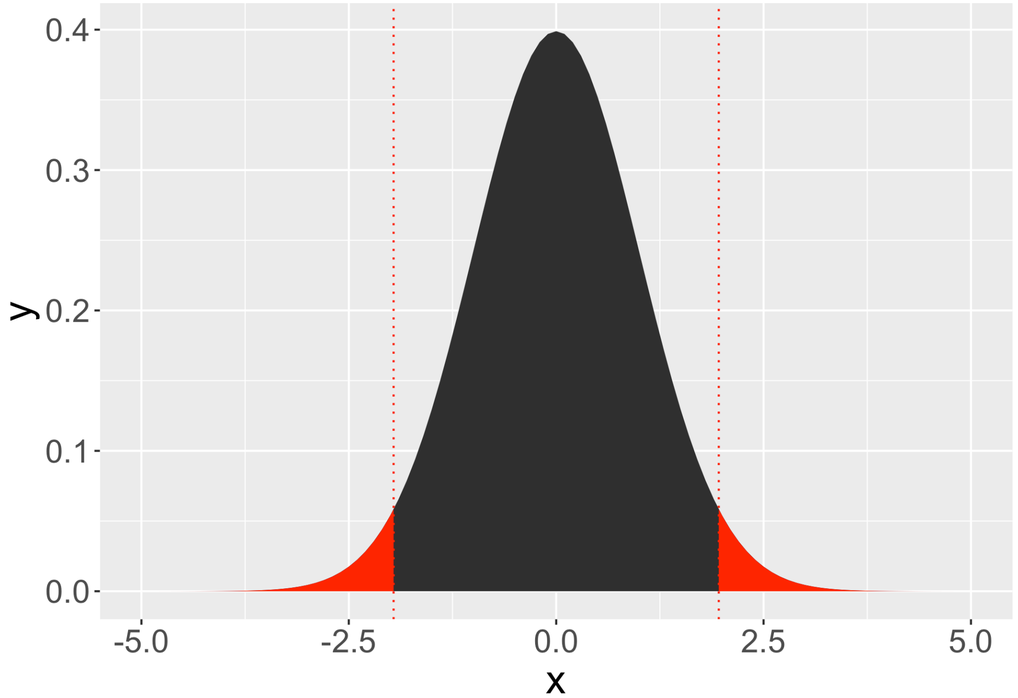
- 1-sided (1-tailed): If we are expecting mean of female to be LESS than mean of females, then our 5% or less can remain at the lower tail
- 2-sided (2-tailed): If we don’t know the direction, then our 5% needs to get split up between both tails (2.5% on either side)

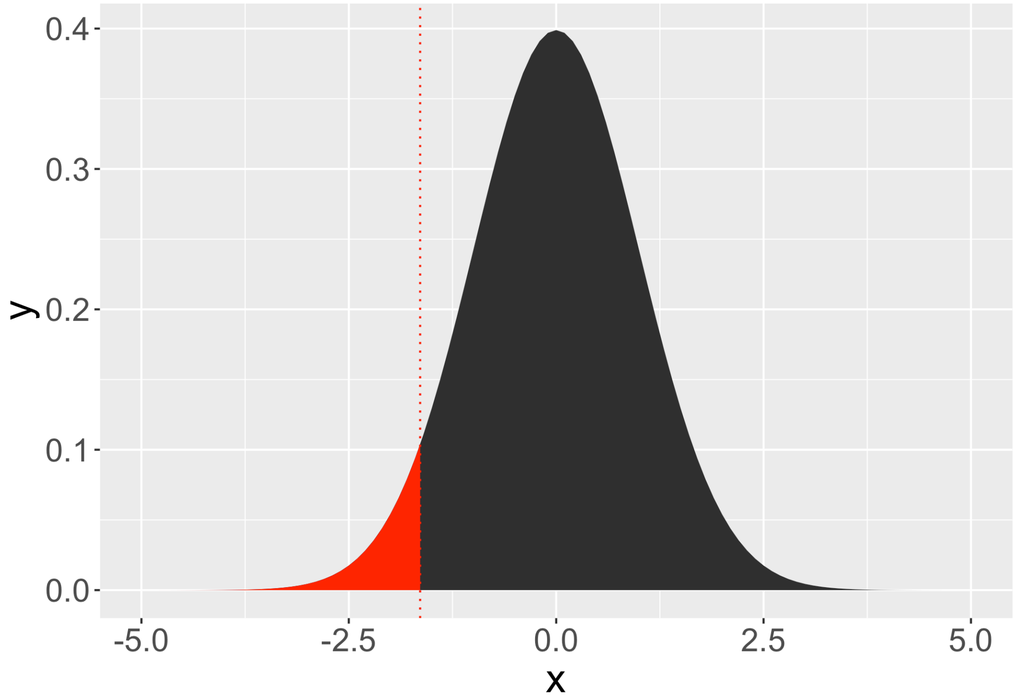
Example

Step 2: Set α

- These cutoffs will be different based on the parameters of the normal that we define (e.g., mean & SEM)
- For standard normal, 1-tailed = 1.64 & 2-tailed = 1.96

$$\text{Critical Value} = \mu_0 + Z_{.95}\frac{\sigma}{\sqrt{N}}$$





Example

Step 3: Collect Data

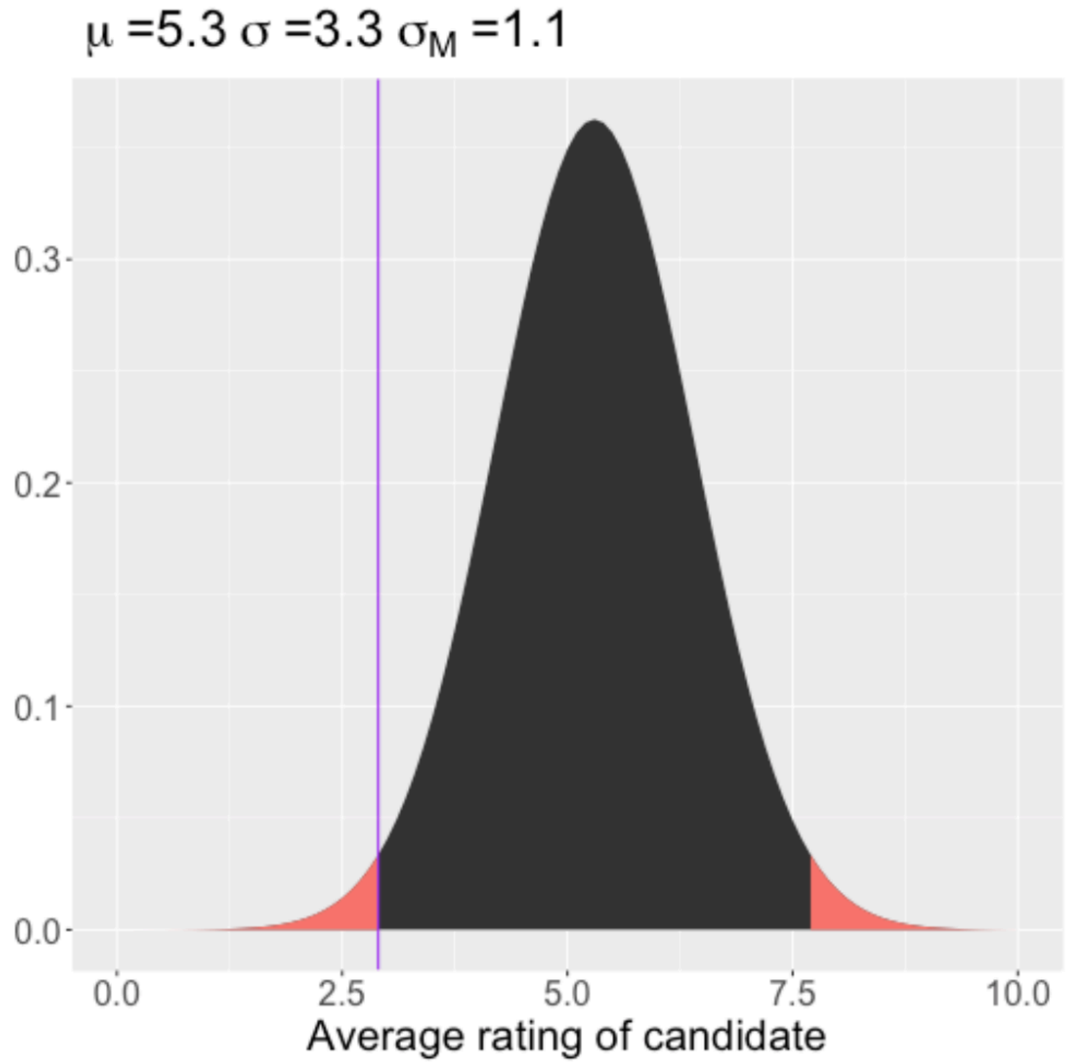
Example

Step 4: Define Sampling Distribution

The mean of the sampling distribution = the mean of the null hypothesis

The standard deviation of the sampling distribution:

$$SEM = \frac{\sigma}{\sqrt{N}}$$



Example

Step 5: Get Probability

- Now we have:
  - A normal distribution, for which we know the mean (μ_M) & standard deviation (SEM)
  - We also have a score of interest, our sample mean (and we want to compare this to our sampling distribution of means)
- Let's use these to get a z-statistic

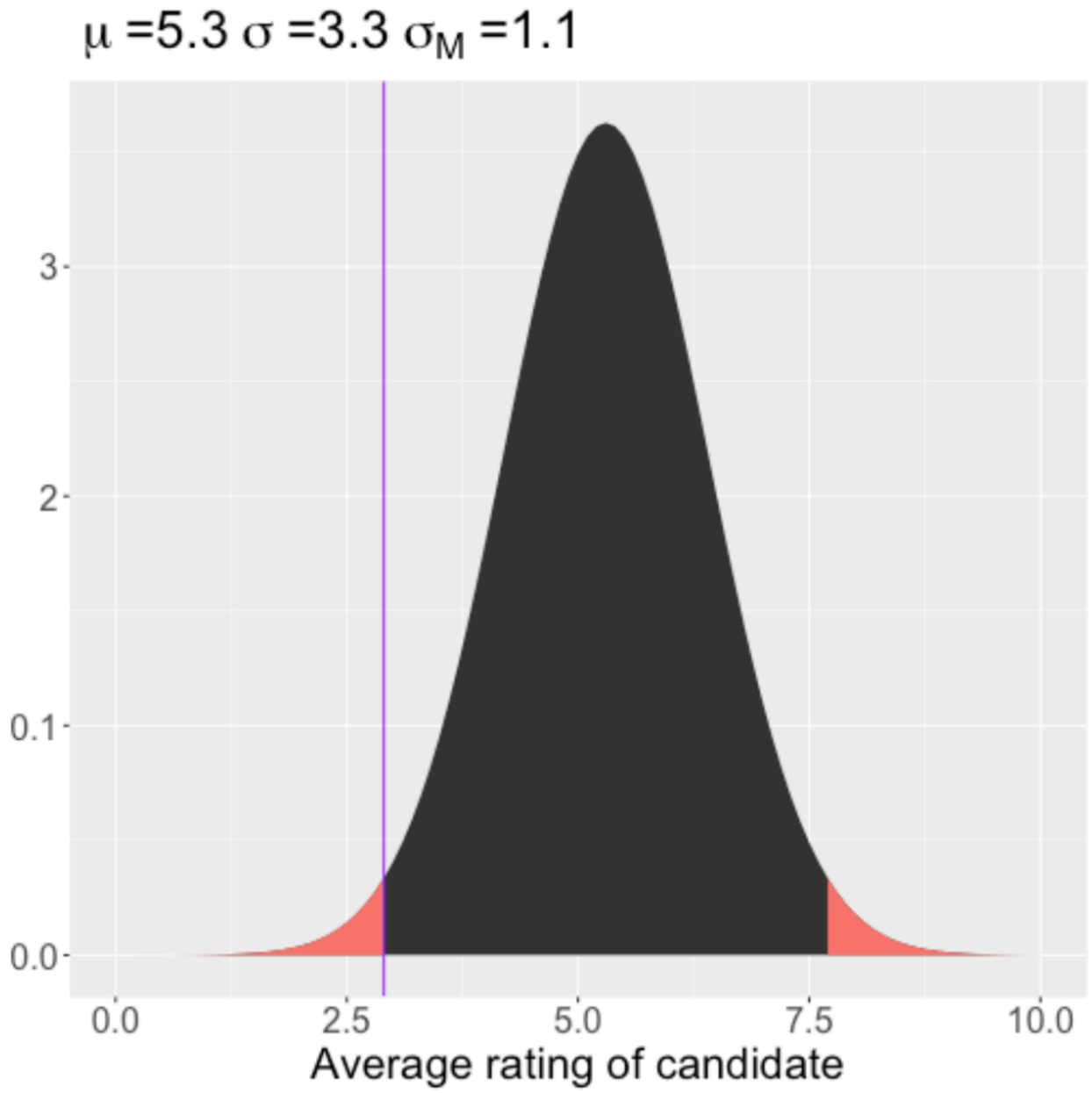
$$Z = \frac{\bar{X} - \mu}{SEM} = \frac{2.9 - 5.3}{1.1} = -2.18$$

Example

Step 5: Get Probability

```
> pnorm(q = -2.18) * 2
[1] 0.02925746
```

The probability that the average female applicant's score would be at least 2.18 units away from the average male score is 0.03.



Example

Step 6: Decision Time

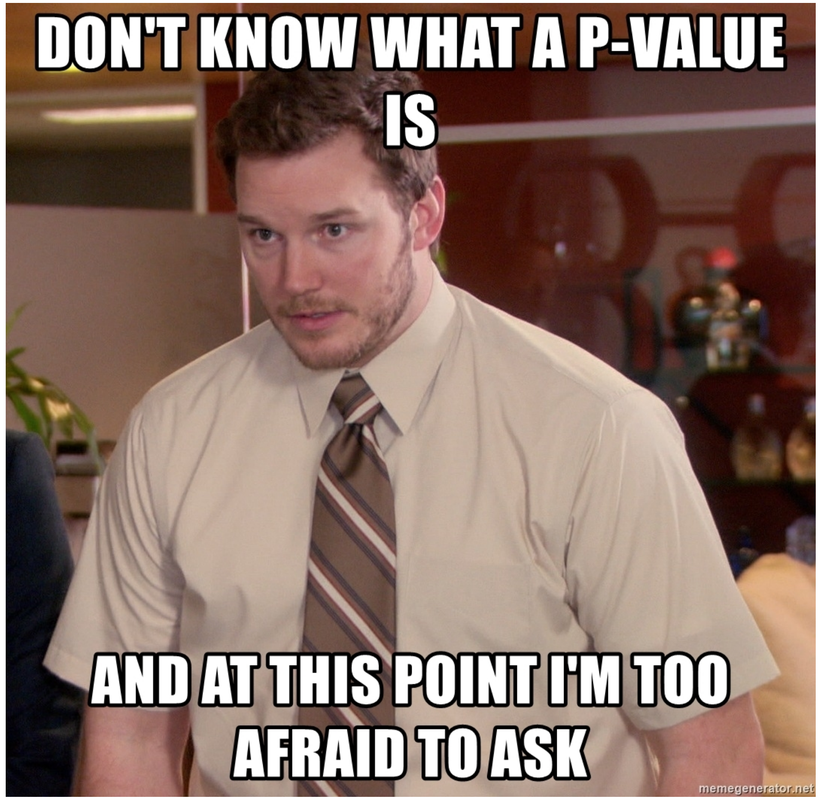
- $p = .03$
- Our p is less than our cutoff of .05
- “statistical significance”
- The mean rating of female applicants is significantly different from the mean of male applicants, $p = .03$

Statistical Tests

- All parametric statistical tests follow this framework
- It's not always with a normal distribution, though. There are other probability distributions! χ^2, *F*, *t*, binomial, poisson, and so many more!
- We use these theoretical distributions to make probability statements (this is the point of parametric tests…that we can use these theoretical distributions)

p-values

- $p(D \mid H_0)$, not $p(H_0 \mid D)$
- It does not…
  - Tell you the probability that the null hypothesis is true
  - ***Prove*** that the alternative hypothesis is true
  - Tell you anything about the ***size or magnitude*** of any observed difference in your data
- What's the difference between p = .04 & p = .06?



How did we get here? — History

No one wanted this

- Fisher — significance tests
  - “Significant” = “worth of further investigation”
  - No decision
  - Evidence against the null
  - Quantify evidence
  - Continuous *p*-values
  - Pulled the .05 rule out of thin air
- Neyman/Pearson — hypothesis tests
  - Acceptance procedure (lots of planning pre-experiment on costs/
  - Binary decision theory
  - Only a decision

https://link.springer.com/chapter/10.1007/164_2019_286

What do we learn from $p < .05$?

As used in the current NHST framework, not classic

- The probability of our alternative hypothesis being correct? **NOPE**
- The likelihood that our theory is correct? **NOPE**
- Is it something we should pay attention to?

NOPE — you need to decide this in the broader context of your theory, the literature, the specifics of your study etc.!

NHST is a decision making tool.
*What if you make the **wrong** decision?*

Errors

- Falsely rejecting the null hypothesis is a Type I error. Traditionally this has been viewed as particularly important to control at a low level (akin to avoiding false conviction of an innocent defendant).

	Reject H_0	Do not reject
H_0 True	Type I Error	Correct decision
H_0 False	Correct decision	Type II Error

Errors

- Failing to reject the null hypothesis when it is false is a Type II error. This is sometimes viewed as a failure in signal detection.

	Reject H_0	Do not reject
H_0 True	Type I Error	Correct decision
H_0 False	Correct decision	Type II Error

Errors

- Null hypothesis testing is designed to make it easy to control Type I errors. We set a minimum proportion of such errors that we would be willing to tolerate in the long run. This is the significance level (α). By tradition this is no greater than .05.

	Reject H_0	Do not reject
H_0 True	Type I Error	Correct decision
H_0 False	Correct decision	Type II Error

Errors

- Controlling Type II errors is more challenging because it depends on several factors. But, we DO want to control these errors. A Type II error is a failure to detect a signal that is present. **Power is the probability of correctly rejecting a false null hypothesis.**

	Reject H_0	Do not reject
H_0 True	Type I Error	Correct decision
H_0 False	Correct decision	Type II Error

Power Analysis

Controlling Type II Errors

- Four quantities are interrelated:
  1. Sample Size
  2. Effect Size
  3. Significance levels (α)
  4. Power
- We must specify a specific value for the alternative hypothesis to estimate and control Type II errors.

Power

Your ability to detect an effect ***if it's actually there***

- What happens if you are underpowered?

NHST

“Good” Science?

- $p < .05$ 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 neuroscience & psychological research
- Unavoidable role of human motives: fame, recognition, ego

NHST

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 sizes in published work
- Underestimation of complexity (why did the failures occur?)
- Underestimation of power
- Inability to replicate
- Value alternative hypotheses/crappy theories

What kind of science have we produced?

- Dichotomous thinking (based on p): research either “succeeds” or “fails” to find 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 (“difference” is enough)
  - No non-nil null hypotheses
  - Weak, slow progress as a science

- "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)

- “… 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'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)

False Positive Psychology

Imagine rolling a die

- Probability you roll a 2?
  - P(2) = 1/6 = 16.7%
- If you roll the die twice, what is the probability that you get a 2 at least once? 30.6%
- If you roll the die 5 times, what is the probability that you get at 2 at least once? 59.8%
- Roll the die enough times, and you'll eventually get a 2. NHST when the null is true is like rolling a 20-sided die.

False Positive Psychology

Simmons et al. (2011)

- But each study is NOT a single roll of the die!
- Instead, each study (even those with only 1 NHST test) might represent many rolls of the die
- **Research degrees of freedom.** Decisions that a researcher makes that changes the statistical test
  - Additional dependent variables
  - Tests with or without covariates
  - Data peeking

False Positive Psychology

Simmons et al. (2011)

- Each time we see how a decisions affected our result, we are rolling the dice again

Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
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%

Just replicate the study…

Duh

Inability to replicate published research

What changes can we make?

- Effect size estimates and CIs, but not p

p-values, CIs, and Effect Sizes

p-values

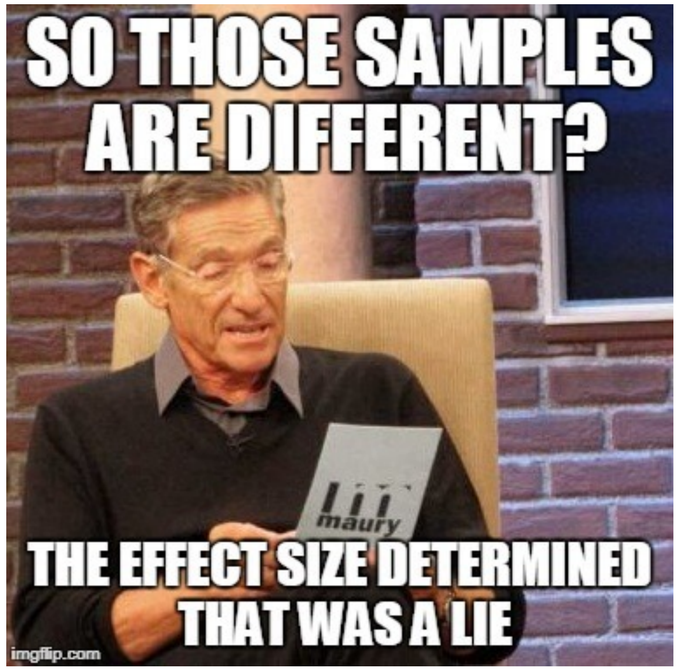
- The probability of getting the observed or a more extreme value of the test statistic, given that the null hypothesis is true
- Binary decision “significant” or “not significant”

Confidence Intervals

- Precision of your estimate
- Wider = worse
- Range gets smaller as N increases
- Most basic is standard error of the mean (67% CI for population mean)

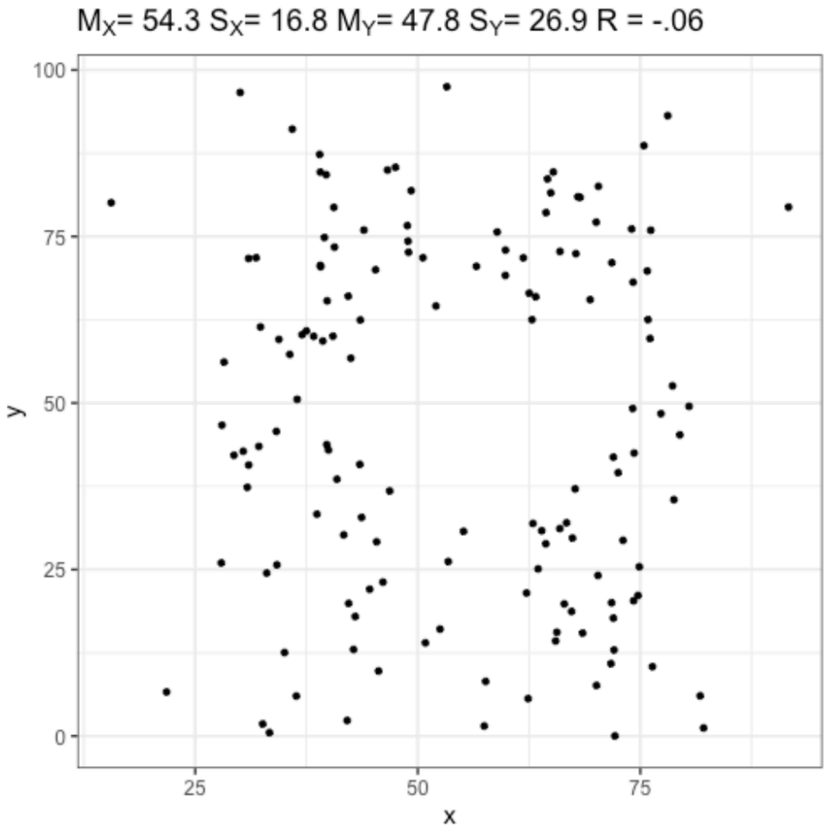
Effect sizes

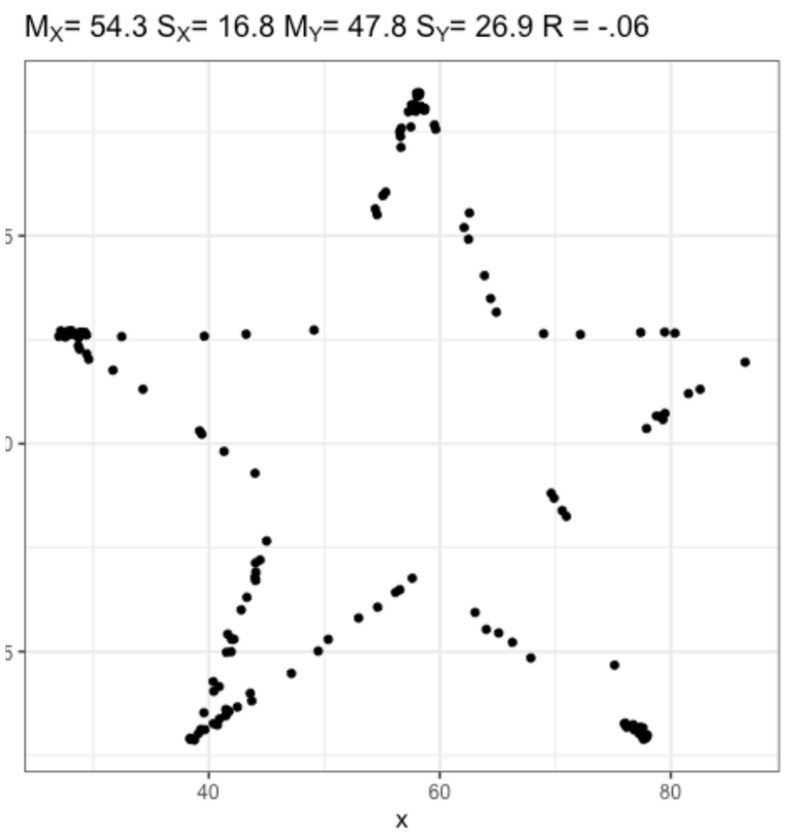
- Magnitude of the effect; “how big?”
- A million effect sizes; most basic is a Pearson correlation
- NOT systematically altered by N

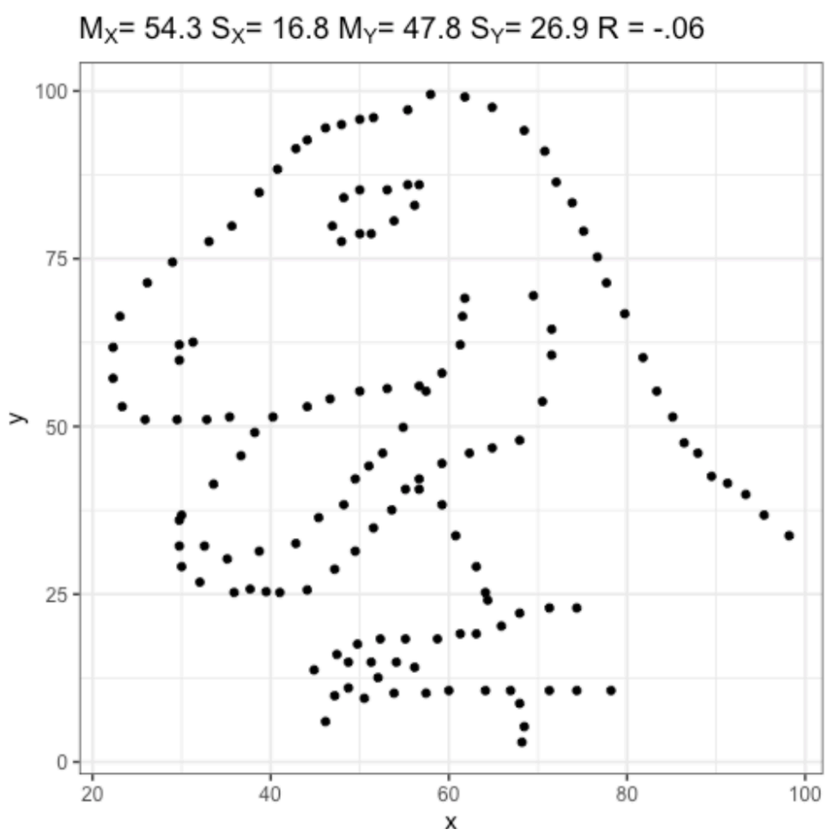


Estimates of effects size and how they were calculated

- Correlation (Pearson's *r*) is the most basic of these







Inability to replicate published research

What changes can we make?

- Effect size estimates and CIs, 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

Challenges to be addressed

- Institutionalization of different procedures…it's hard to move away from a seemingly helpful heuristic!
- Journals need to change their values (why should they, they are making money)
- How do we shift media values? Replication is not sexy
- How do we shift academic values? What will impact tenure? How do we re-calibrate "productivity"? Citation indices?

Is Psych bullshit?

- Inability to replicate has been viewed as a singular failure of psychology as a science
- But…
  - Literally every field is struggling with this — it's not just psychology
  - At minimum, it tells us that we don't know things as well as we think we do. Is that a failure of science?
  - Can we progress?

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*

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

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
- More sophisticated methods of inference also exist. The rest of this class is is exploring what those frameworks look like.
    - As we get further along, you will a move away from p-values and towards fit measures. These have room for interpretation, but give a more holistic picture. **It is an art form, not a science!**