

BIOMED SCI 552:

STATISTICAL THINKING

LECTURE 8: HYPOTHESIS TESTING

QUESTIONS FROM TUESDAY?



THE FINAL

- From the Syllabus: “You will select a topic of interest to you, based on your own work or as part of the material we have covered over the course of the semester. The purpose of this project is to identify a problem – including prior research in the area, and outline how a statistical approach may aid you in answering it.”
- What does this mean?
 - This is your chance to apply what you’ve learned in this class to a problem you’re interested in
 - It should be biomedical, but it need not be the project you are currently actively working on, your dissertation topic, etc.
 - Essentially, what I’m looking for is a brief review of the literature to introduce the topic, followed by you considering the statistical approach to your problem up until you have to *do* the statistics
 - What is your source population? Do you have a target population? A sampling frame? What considerations would you put into sampling? Are there major threats to your approach – bias in sampling? Logistical difficulties? Possibly nonpositivity?
 - What type of hypothesis are you testing, or is this “just” estimation
 - Not hypothesis testing and “What’s your hypothesis” in the colloquial sense can be different

FORMATTING AND OTHER GUIDELINES

- Your final paper should be 5 pages
- This is a maximum, not a minimum
- I'm not going assign specific formatting guidelines, but don't be clever
- A broadly applicable piece of advice in academia for grantwriting, getting papers through peer review, etc.: Don't annoy your reviewers
- CITEYOUR SOURCES
 - Again, there is no specific citation format, so pick one

WHAT IS HYPOTHESIS TESTING?

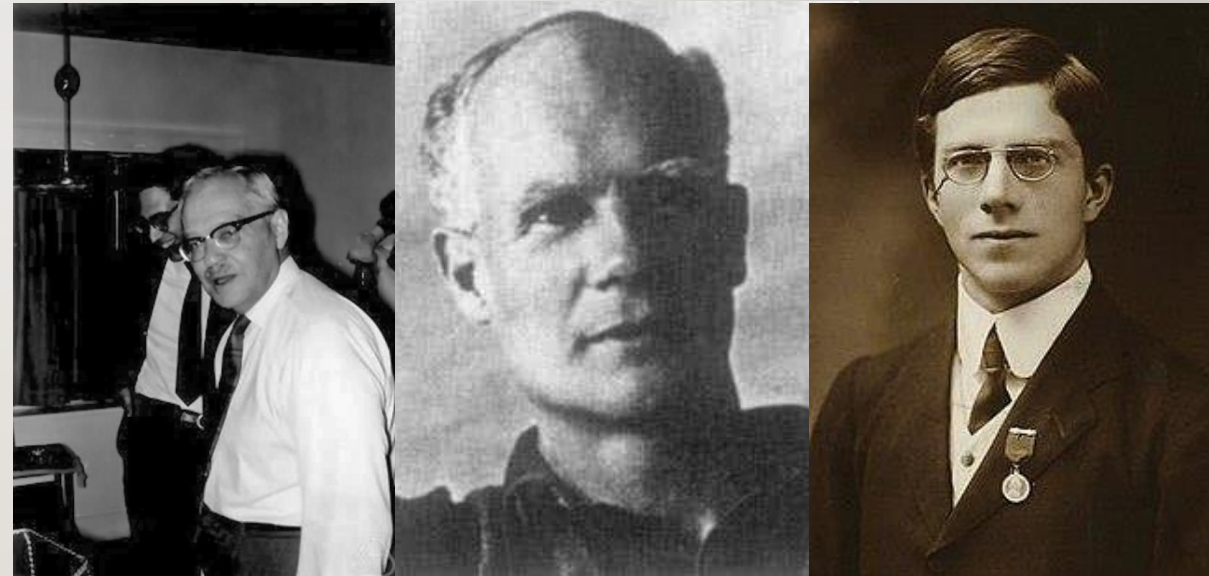
- Hypothesis testing is a formal procedure for interrogating how well our ideas about the real world cohere with with our data
- It is the process of determining how well a “null” hypothesis about a *population* quantity fits a *sample* of the data
 - This bit is important – while we are using our sample data, we are asking a question about the population
- Estimation can tell you how big an effect is, while hypothesis testing tells you if there’s any effect at all

WHAT IS A *STATISTICAL* HYPOTHESIS?

- A statistical hypothesis is a precise, quantitative statement about a population, which is then evaluated to see if the sampled data supports it
- In hypothesis testing, we first consider a “null hypothesis”
 - This is *often* but need not always be a “nil hypothesis” (i.e. a statement that there is no effect)
 - Usually, this is denoted H_0
- There is also an “alternative hypothesis”
 - This is usually denoted H_A
 - This is a mutually exclusive alternative to the null hypothesis

WHERE THIS GETS MESSY AND COMPLICATED

- Hypothesis testing is often called “Neyman-Pearson Hypothesis Testing” after two collaborators who developed much of the theory
- Ronald Fischer *also* developed much of the theory, and popularized the use of the phrase “significance test”
- They disagreed on whether or not you needed an alternative hypothesis
- This debate was not settled when Fisher died in 1962



Neyman, Pearson and Fisher

FISHER'S APPROACH

1. Set up your null hypothesis
2. Report the level of significance you find (we'll discuss what that means soon)
3. If your result is not significant, you can draw no conclusions or make no decisions without more data

Fisher advised only using this when little is known about your problem – he started life as a Bayesian, and asserted that these were sort of provisional findings that ignored wider data.

This is sometimes called “Null Hypothesis Testing” for obvious reasons, and is often most concerned with asking if the observed data are consistent with having arisen from chance

NEYMAN-PEARSON APPROACH

- Set up a null and alternative hypothesis, and determine acceptable rates of alpha (the rate of false positives) and beta (the rate of false negatives) based on considerations of costs-benefit, etc.
 - Even relatively cold hypothesis testing involves subjectivity
- If the data falls into the rejection region of the null hypothesis, accept the alternative hypothesis
 - This does not mean you think it is true, just that you'll act on it as if it is
- Here you *need* disjoint hypotheses, and you need to be able to meaningfully make a cost-benefit argument for alpha and beta

THE MODERN METHOD

- We just sort of...crash the two into each other
 - We start using Fisher's p-value concept as a substitute for Neyman-Pearson's more complicated likelihood ratio test in many cases
- This is a messy hybrid that has a whole mess of hidden assumptions and caveats
- Neyman and Pearson arguably had the much more rigorous form, but what is often taught is much like Fisher's conception
 - This includes things like primarily being concerned with an alternative hypothesis that is "Not The Null Hypothesis" rather than an actual alternative hypothesis
 - Neyman-Pearson testing allows for the estimation of both Type I and Type II error
 - But we've largely abdicated any meaningful discussion about our choices of hypothesis and alpha and beta to field-specific conventions

TYPE I AND TYPE II ERROR

- Type I error: This type of error occurs when we reject a null hypothesis that is actually true
- Type II error: We fail to reject a null hypothesis that is actually false
- Because we are often looking for an effect in a population, these are referred to as “False Positive” (i.e. we said an effect existed when there wasn’t one) and “False Negative” (we said there wasn’t an effect when there was) but conceptually, they can be broader than that

IS THIS A GOOD IDEA?

- “We are quite in danger of sending highly trained and highly intelligent young men out into the world with tables of erroneous numbers under their arms, and with a dense fog in the place where their brains ought to be. In this century, of course, they will be working on guided missiles and advising the medical profession on the control of disease, and there is no limit to the extent to which they could impede every sort of national effort.” – R.A. Fisher
- Fisher was an experimentalist and evolutionary biologist in addition to being a statistician, and was concerned (rightly) with how very clean mathematical notions of probability intersected with the realities of empirical science
- Before we get too carried away, he was also a eugenicist

THE END RESULT

- The end result of this is confused students (as predicted)
- Hypothesis testing is massively powerful, often quite misunderstood, and in practice we somewhat limit ourselves to well understood hypothesis tests in many cases
- The *bulk* of hypothesis testing involves a null hypothesis, and an alternative hypothesis that is simply “not the null”
- Quite often, H_0 is that a particular sample distribution is indistinguishable from one arising from chance, or for the estimates of many effects, that the effect is zero
- This is not always the case. Likelihood ratio tests crop up in model selection, which are testing disjoint hypotheses that one model is better supported than another

LADY TASTING TEA

- Many of you may know that the British have particularly strong feelings about the combination of milk and tea
- Namely, that you add milk to tea, not the other way around
- Does anyone know *why*?

LADY TASTING TEA

- Dr. Blanche Muriel Bristol, who worked with Fisher at the Rothamsted Experimental Station, claimed to be able to distinguish the taste of tea added to milk vs. milk added to tea
- Fisher thought this was nonsense
- A third colleague, William Roach, suggested they test her
 - She would, incidentally, marry William four years later
- They devised an experiment where four cups of each type were presented to her, randomly

WHAT'S THE NULL HYPOTHESIS?




THE EXPERIMENTAL SETUP

- There are 70 possible combinations of results:

Successes	Combinations	Number of Combinations*
0	XXXX	1
1	XXX√, XX√X, X√XX, √XXX	16
2	XX√√, X√X√, X√√X, √X√X, √√XX, √XX√	36
3	X√√√, √X√√, √√X√, √√√X	16
4	√√√√	1

*This follows what's called a hypergeometric distribution



HOW DO WE EVALUATE THE NULL HYPOTHESIS?

- Fisher thought that a 1-in-20 chance of mistakenly rejecting a null hypothesis that was true was acceptable
- Putting it in Neyman-Pearson terms, $\alpha = 0.05$
- This turns into a p-value of 0.05 as a cutoff
- Would anyone like to hazard a guess as to how he determined this?



BACK TO TEA...

- Out of the 70 possible combinations, only one combination (all four cups correctly identified) had a less than 5% chance of occurring by chance (having a 1.4% chance)
- This meant that Fisher would only reject the null hypothesis that Dr. Bristol could not detect a difference if she correctly identified all four cups
- She did so



THE IMPORTANCE OF ASKING THE RIGHT QUESTION

- It should be obvious now that the details of your null hypothesis is critically important
- “Greater than”, “Less than”, and “Different from” mean different things, and have mathematical consequences
- You want to be sure your statistical hypothesis is in support of the scientific question you think you’re asking
- Importantly, you want to know *what your statistical hypothesis is*
 - ”Here’s my data, I need a p-value” will kill the soul of any statistician you work with

QUANTIFYING UNCERTAINTY

- All of this is about quantifying uncertainty
- How much does our data support a particular argument about something that cannot be measured in the population, but can be measured in our sample?
- We have tools to assess this:
 - Estimation: Confidence Intervals
 - Hypothesis Testing: P-values
 - Bayesian Methods: Posterior Distributions
- The first two are two sides of the same coin
- The third we'll discuss in a later class

SIGNIFICANCE LEVEL

- Looping back, your significance value (alpha) is the probability threshold that defines the binary decision as to whether or not something is “statistically significant”

$$\alpha = \Pr(\textit{reject Null} \mid \textit{Null} = \textit{True})$$

- We say a result is statistically significant when $P < \alpha$
- Alpha is fixed to the amount of uncertainty we're okay with
 - This is most often 0.05 – though as mentioned previously, this is largely arbitrary
 - For some fields, it's different. GWAS studies, for example, demand much smaller values for alpha
- Your confidence level is $1 - \alpha$
 - Hence 95% confidence intervals

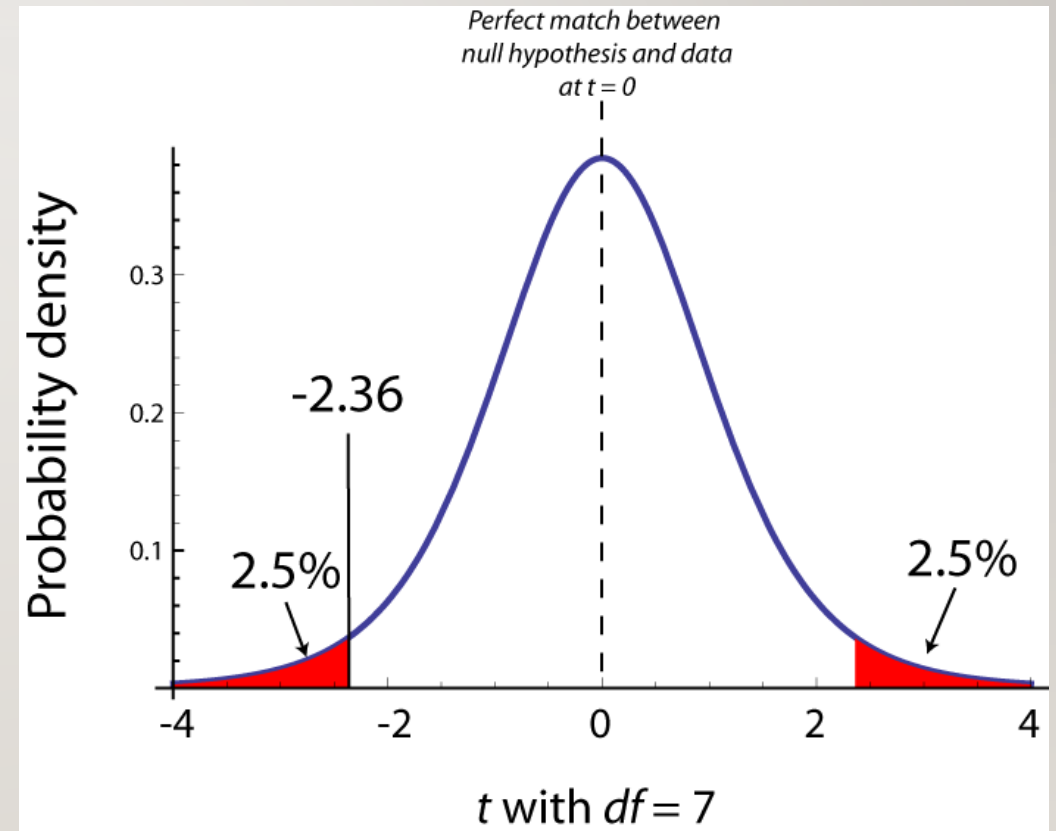
P-VALUES

- Think stochastically – your data is one of *many* potential random draws from your population
- The p-value, which we calculate from a statistical test, describes how likely your particular data would have arisen if the null hypothesis was true
 - i.e. $\text{Pr}(\text{Data} \mid \text{Null Hypothesis})$
- A p-value of 0.17 means there's a 17% chance your data could have occurred under the null hypothesis
- If $p < 0.05$, there is a less than 5% chance
- This is *not* the probability that the alternative hypothesis is “true”, the probability you’re right, etc.

CRITICAL VALUES AND THE NULL DISTRIBUTION

- Null distributions are used to quantify the uncertainty of your hypothesis test.
- Null distributions are classically the distribution of a test statistic when the null hypothesis is true.
- Student's *t*-distribution:

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$



ON SIGNIFICANCE

- Significance. Is Arbitrary.
- Yet we treat it like it's very, very important
- The hurdle between getting a study published when $p=0.051$ and $p=0.049$ is massive, and yet if I told you I had a drug that could cure cancer, but it wasn't significant at $p=0.051$, you'd still be interested
- Don't fall into the rote tendency to just check if $p<0.05$ and treat that as the sole arbiter of if something is interesting or not

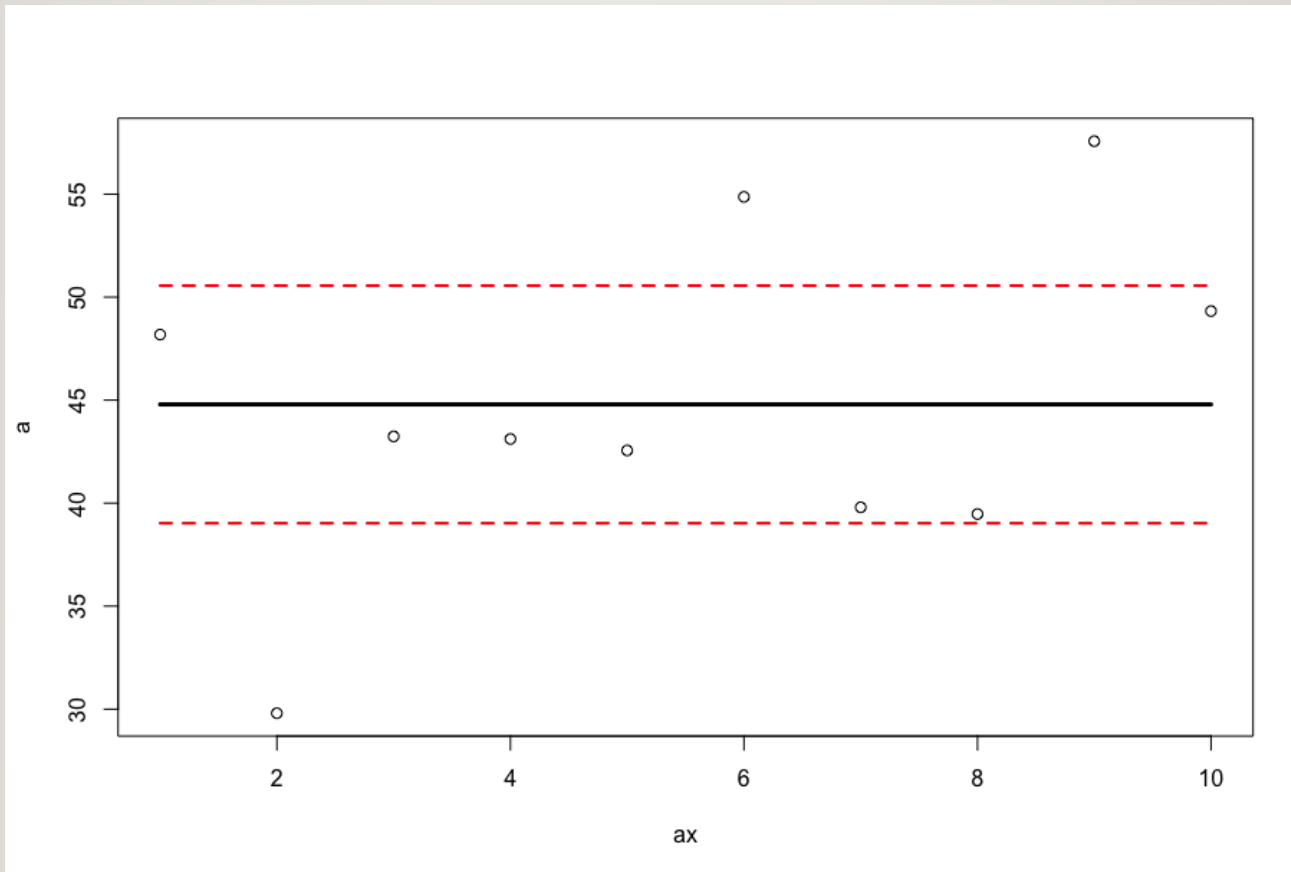
CONFIDENCE INTERVALS

- Confidence intervals are how we measure the uncertainty of estimates
 - This can be a proportion, a variance, the difference between two population means, a correlation coefficient, etc.
- All estimates have sampling distributions and standard errors
- These are most often already determined for us and baked into our statistical tools
- Typically, a 95% confidence interval is $\text{your estimate} \pm 1.96 * \text{standard error}$
- This is the range under which, under repeated sampling, 95% of your estimates would fall

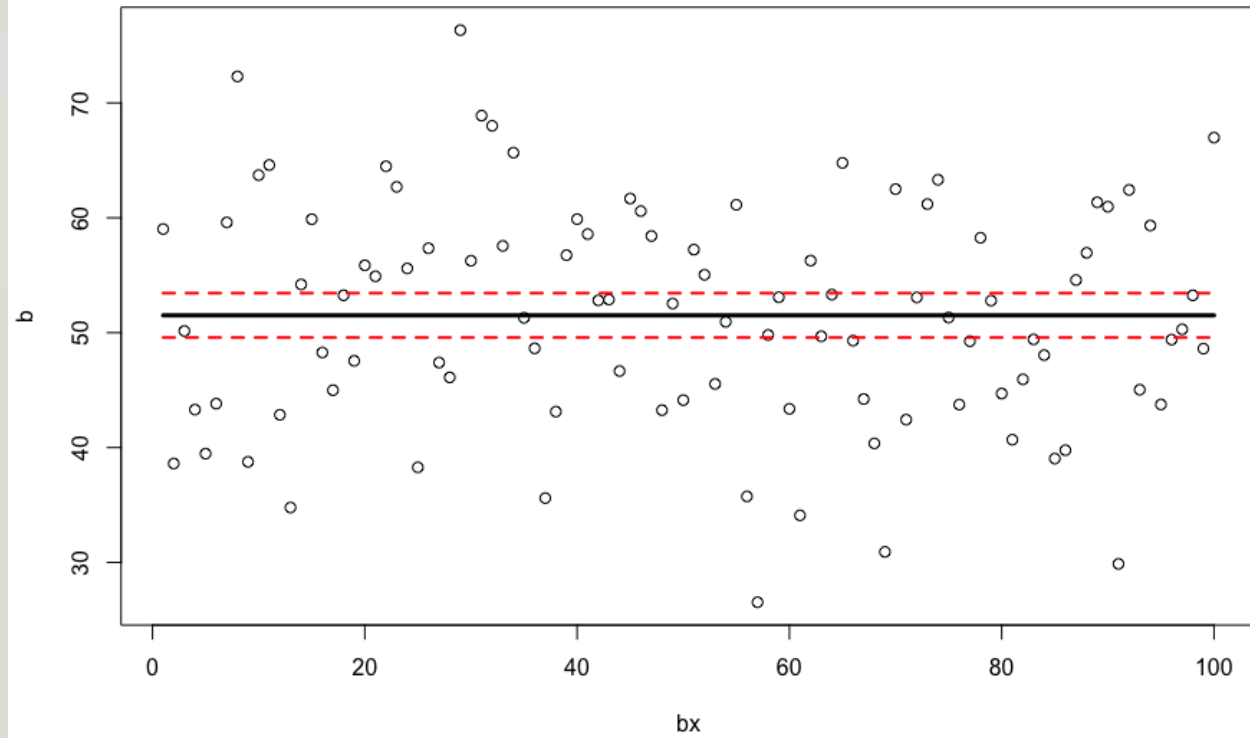
REMINDER...

- A p-value or confidence interval is a measure of *uncertainty*, not of variability
- We become more certain, for example, of our estimate
- But there may still be variability, and that we're just fleshing out if we, for example, increase our sample size

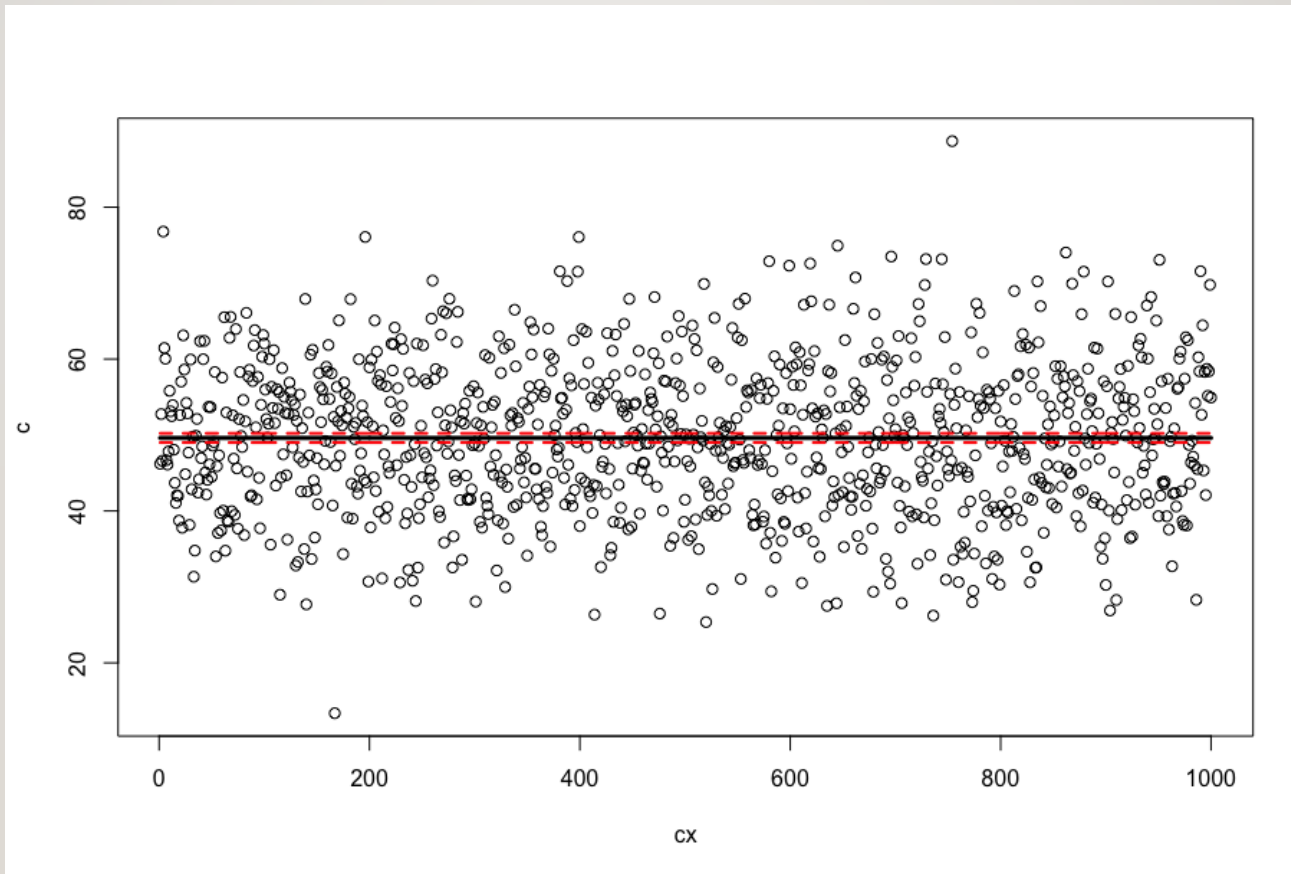
10 SAMPLES



100 SAMPLES



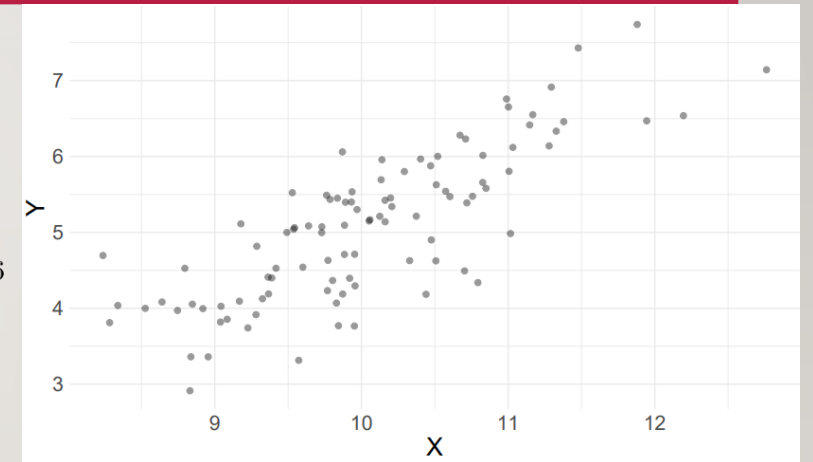
1000 SAMPLES



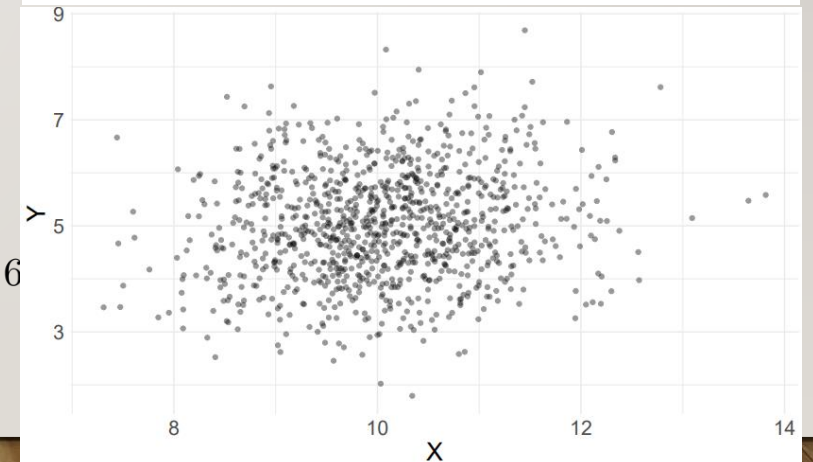
EFFECT SIZE VS. STATISTICAL SIGNIFICANCE

- Statistical significance tells you about your data as compared to some null hypothesis
- It does not tell you if an effect size (correlation coefficient, different in means, etc.) is *important* in the real world

$$r = 0.80$$
$$[0.72, 0.86]$$
$$P < 2.2 \times 10^{-16}$$



$$r = 0.14$$
$$[0.080, 0.20]$$
$$P < 7.296 \times 10^{-6}$$



Courtesy of Erin Clancey

SIGNIFICANT \neq IMPORTANT

- Statistical significance tells us how confident we can be in rejecting the null hypothesis
- What it doesn't tell us if that is important in a clinical, biological, ecological, etc. sense
- What is important depends on *context*
- It's good practice to present as much information as possible
 - Don't just say "Significant or Not"
 - Avoid "Significance Stars"
 - Present the actual p-value
 - Ideally, present the effect with the confidence interval
 - Some journals, like *Epidemiology*, particularly frown on p-values for some of the problems I've discussed earlier
 - This is also important for reproducibility, meta-analysis, etc.