Intro to R Workshop – Part 2

UBC Databinge Data Champions Team

July 13th, 2022

Why are we doing this workshop?

- Biomedical Researchers and Graduate Students come from a wide variety of training backgrounds.
- Not everyone has had the opportunity to spend time on learning how to do statistics and visualizations using coding languages such as Python and R.
- Course offerings at UBC (like Stats 540, 545) presume a working knowledge of R and statistics or have a very steep learning curve.
- Biomedical research publications rely heavily on statistics including hypothesis testing. Very difficult to interpret results or publish your work without the “right stats”.
- Coding languages like R, when used in conjunction with Github and Open Science Framework, make for computationally reproducible workflows

What did we cover in Part 1?

- Getting Oriented with RStudio.
- The basics of coding in R.
- Introduction to plotting in R using GGplot.
- (very) Brief introduction to statistics in R.

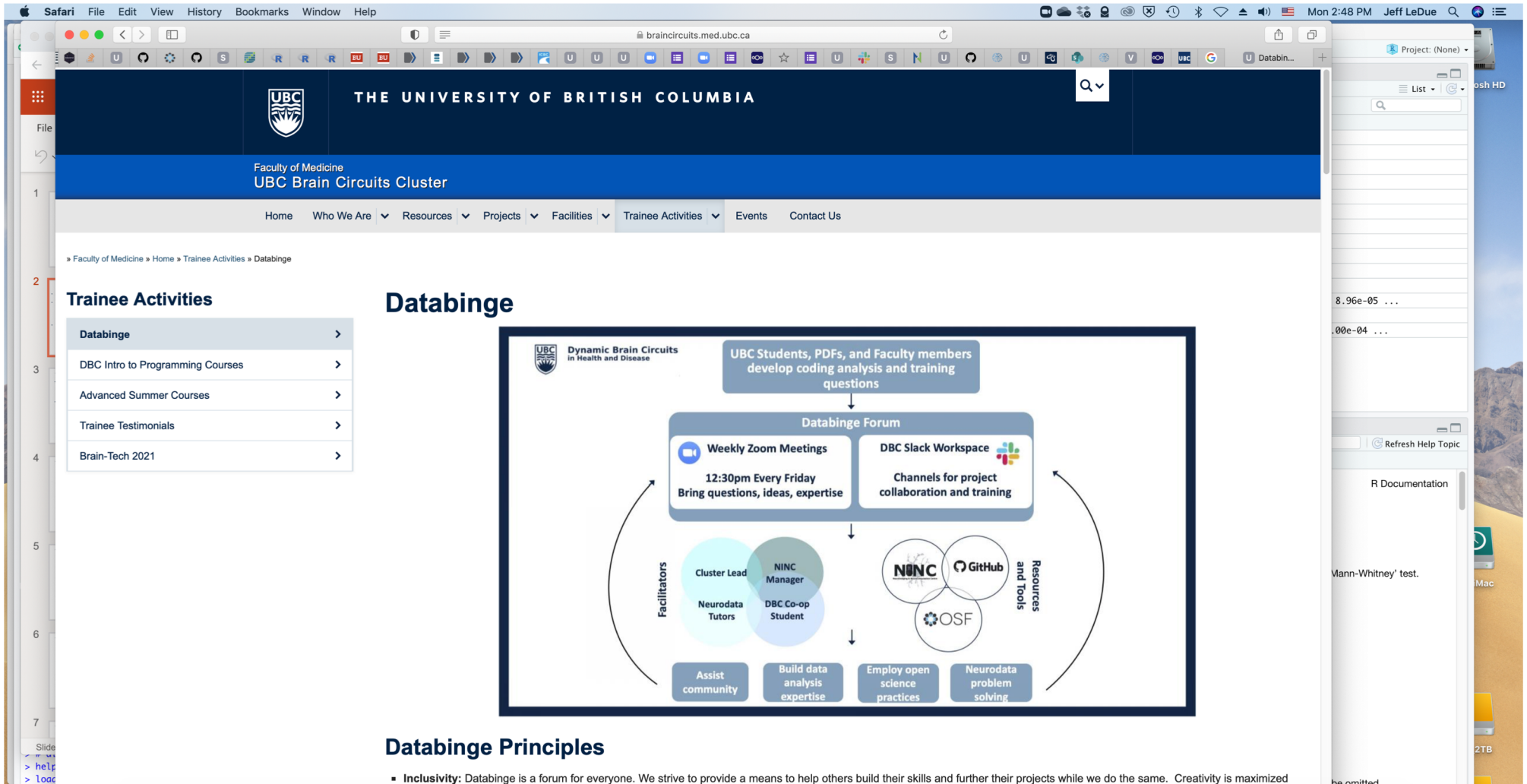
What will we cover in Part 2?

- Demo of a reproducible workflow using OSF and RStudio
- Review t test from Part 1
- Variations of the t test, some theory, and in RStudio
- Power calculations to help with experiment design
- Introduce ANOVA

Learning goals for Part 2:

- Familiarity with using tools like RStudio, Open Science Framework and Github to make a computationally reproducible workflow.
- Increased familiarity with R/RStudio
- Comfortable performing variations of the t-test (e.g. 1-sample, 2-sample, paired) in R.
- Assessing the Null Hypothesis based on the observed value of the sample statistic (i.e. mean), its sampling distribution and its corresponding p-value.
- Distribution free alternatives incl. Wilcoxon Rank Sum test, Permutation testing.
- ANOVA, what is it? how do I run one in RStudio?

Where can I get more support?



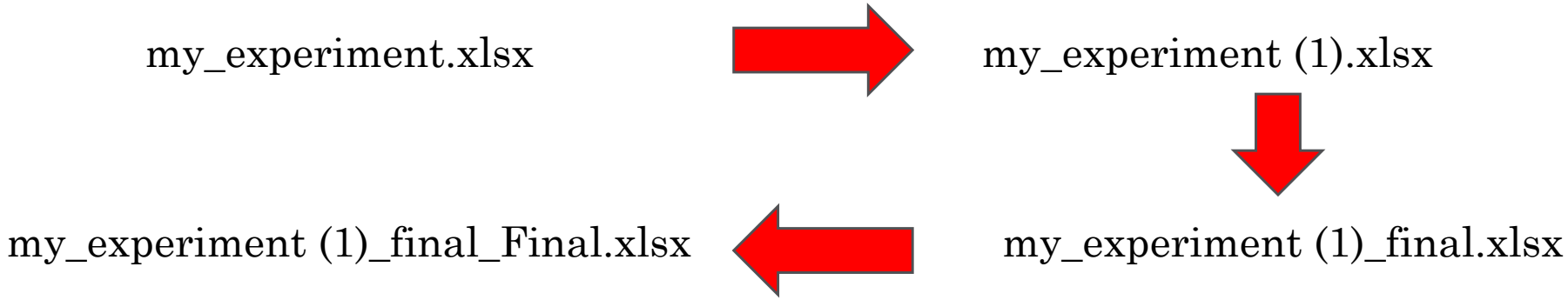
Reproducible workflow walkthrough

UBC Databinge Data Champions Team

July 13th, 2022

Why reproducible workflows?

- Two reasons:
  - Support Open Science by making our work computationally reproducible. This means that your figures or other outputs can be reproduced independently by another researcher by running your script.
  - You will save yourself time and confusion.
- Working with GUI based programs like excel and prism requires a lot of human interaction.
  - Humans don't do well with repetition. We make mistakes.
  - Projects get disorganized very quickly. Will you remember what you did in 6 months? 6 years?



Tools for reproducible workflows

- A coding or scripting language = R/RStudio
- A versioning system for scripts = Github
- A repository for data that allows programmatic access = Open Science Framework
- Well developed OSF project as an example: https://osf.io/h3ec5/
- Test project for our reproducible workflow walkthrough: https://osf.io/k5s4f/

Walkthrough of reproducible workflow

- Main steps:

- Retrieve the script from github: https://github.com/ubcbraincircuits/datachampions-R-part2
- Load in RStudio.
- First steps download the data file from the OSF project (it is the data from part 1)
- Workflow consists of subsetting the data (as in part 1), running stats tests and making plots.
- Plots could then be stored back on OSF and included in your posters, papers and other outputs. This can done with the browser or programmatically.
- You will be certain you can reproduce them in the future and others can too.

Hypothesis Testing

UBC Databinge Data Champions Team

July 13th, 2022

Review of "homework" from Part 1

- In Part 1 we introduced the command in R for a one sample t test
- Two sample t test was left as an exercise
- Let's move over to R and review some of the commands before we go into more theory about hypothesis testing.

A Few Important Definitions.

- Population: all of the individual units of interest (what we want to know about)
- Sample: a subset of the population that is studied (what we can actually observe).
- Descriptive statistics: values calculated from samples that describe the sample. Examples include the mean, median, and standard deviation.
- Inferential statistics: values calculated from samples that let you learn something about the data. One example is a P-value.

Hypothesis Testing.

- Ho : The Null Hypothesis.
- H_A : The Alternative Hypothesis

We collect evidence to see if we can or cannot reject the null hypothesis. We never accept the null or the alternative hypothesis.

The null and alternative hypothesis should be decided before you start the experiment.

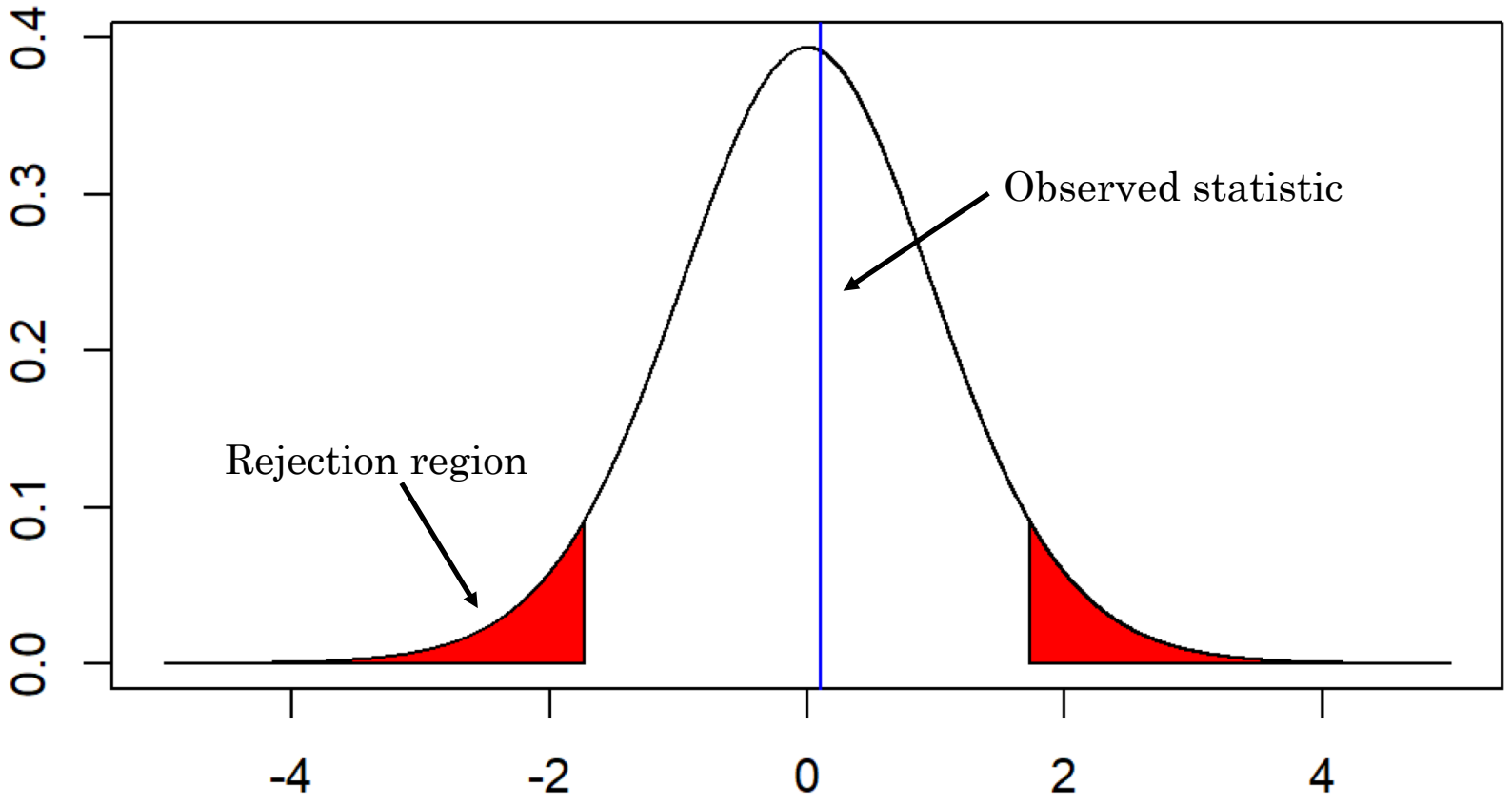
Error in Hypothesis Testing

Reality	What you conclude: Ho CANNOT be rejected	What you conclude: Ho is rejected
Ho True	✓	Type 1 error
Ho False	Type 2 error	✓

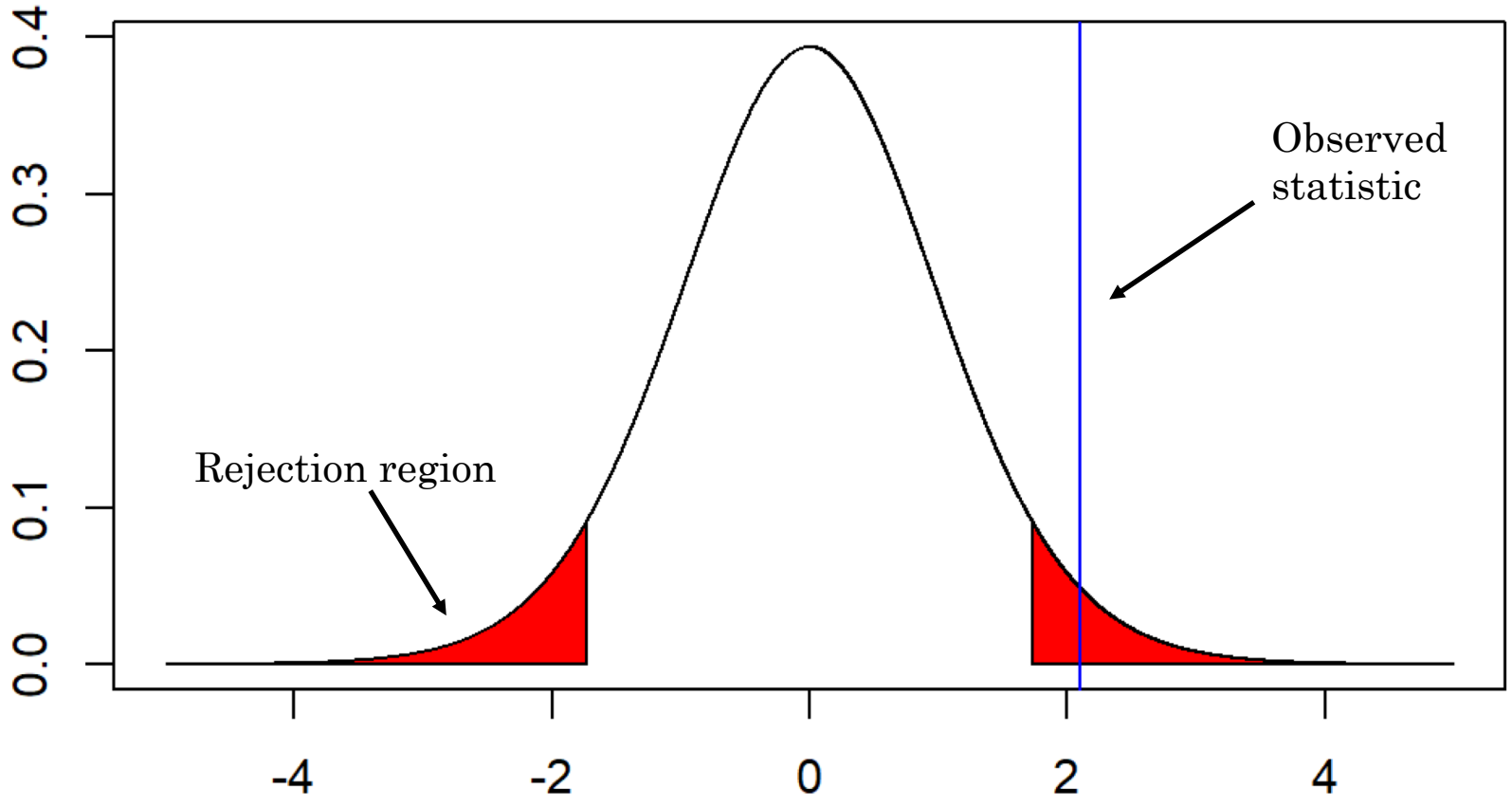
Using Statistics to Test Our Hypothesis

- In order to actually test our hypothesis, we need a few things:
- **A test statistic** that is sensitive to what you are interested in (eg. t statistic for t-test, w statistic for Wilcoxon)
- The **observed value** of the test statistic. (e.g. calculated from your data)
- How that statistic is distributed if the Null Hypothesis is true (e.g. **the sampling distribution**)
- If our observed value of the statistic is in the tail of the sampling distribution this means it is a rare observation and we may be able to reject the Null Hypothesis.
- If our observed value of the test statistic is in the middle of the sampling distribution, we will not be able to reject the Null.

Visual explanation of hypothesis testing



Visual explanation of hypothesis testing



The Student's t-statistic.

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

X = Sample Mean μ = mean under the null hypothesis

s/sqrt(N) = Standard error of the sample

- Two major points to take from this equation
  - The student's t statistic is a measure of the difference between the sample mean and the mean under the null hypothesis.
  - The t statistic changes depending on the sample size.
- Sampling distribution (t-distribution) can be calculated theoretically

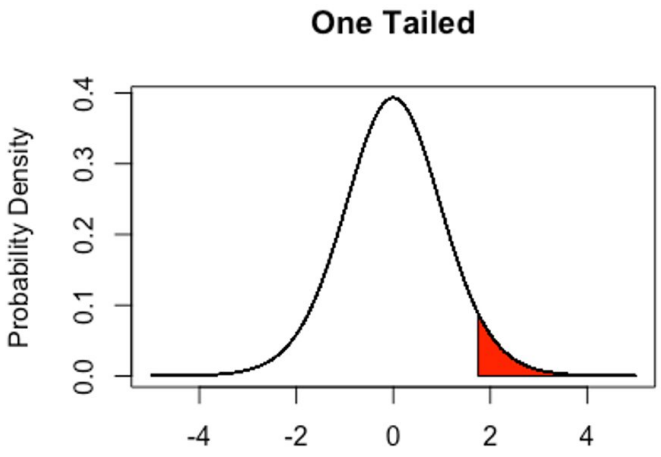
One Tail vs Two Tail.

- One tailed and two tailed tests are referring to the rejection region
- Deciding between these two is entirely a function of how you are wording your alternative hypothesis.
  - Greater than the null?
  - Less than the null?
  - More extreme than the null?

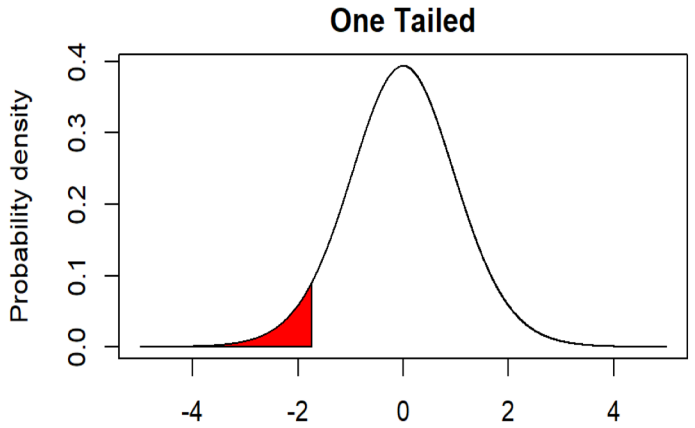
Note: How you word your alternative hypothesis should be based on what results you are interested in.

One Tail vs Two Tail visually.

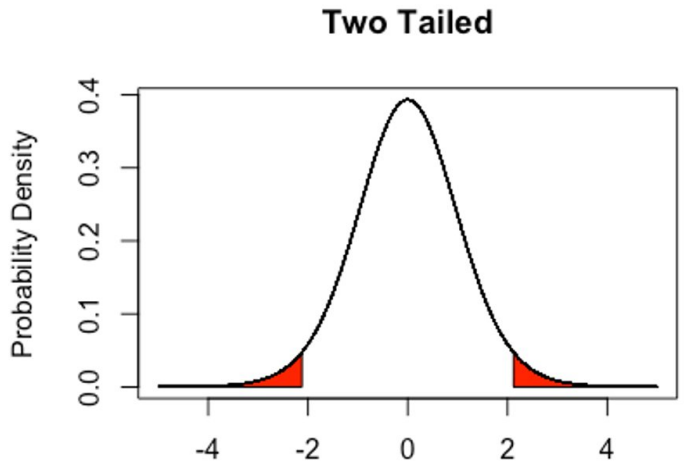
Greater than the null



Less than the null



More extreme than the null



Three Different Kinds of t-tests.

- **One sample t-test**
  - Used to test if the sample mean is different from what we would expect if the null hypothesis was true.
- **Two sample t-test**
  - Used to test if the means of two samples are different from each other.
- **Paired t-test.**
  - Used to test if a population undergoes a significant change after a certain treatment. Essentially you are doing a 1-sample test on the differences before and after the "treatment".

Assumptions of the different t-tests

- One sample t-test
  - Random sampling
  - The variable is normally distributed.
- Two sample t-test
  - Random sampling in both populations
  - The variable is normally distributed in both populations.
  - The standard deviation of the variable is the same in both populations.
- Paired t-test.
  - Random sampling
  - The variable is normally distributed.

When the t-test won't work and what you should do.

- What do we do if our sample statistic does not seem to be normally distributed and our sample size is small?
  - Wilcoxon rank-sum test (aka Mann-Whitney)
  - permutation testing.
- What if we have more than two groups we need to compare?
  - ANOVA and multiple comparisons.

Wilcoxon rank sum test

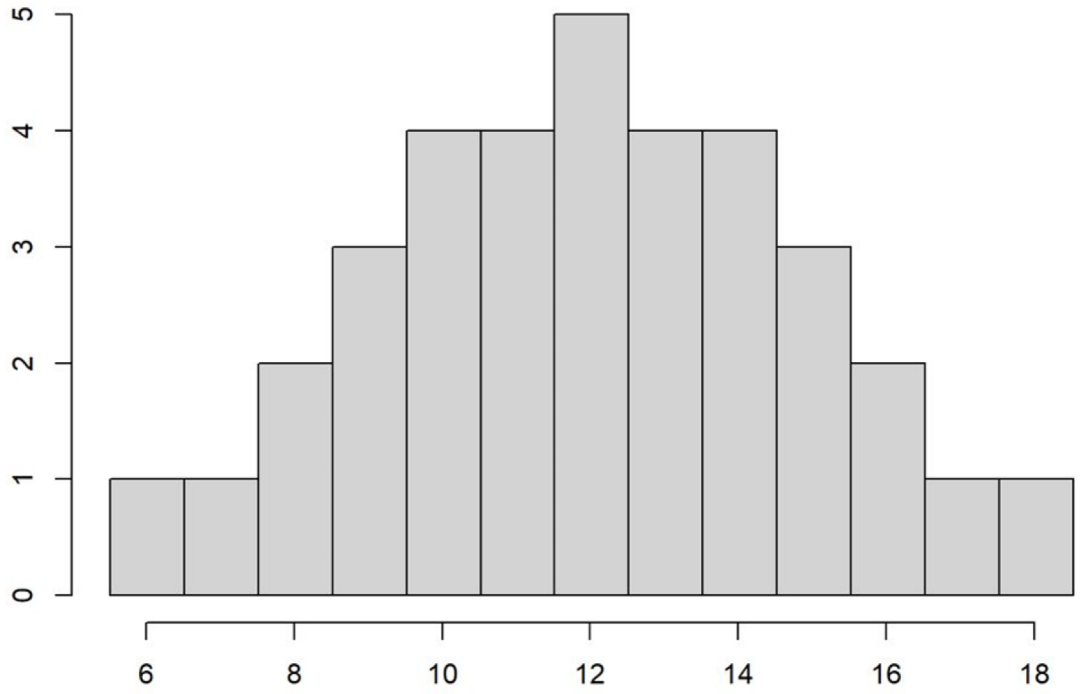
- distribution free alternative to the 2-sample t-test. Can be used for non-normal distributions, in fact does not make any assumption about the shape of the underlying distribution.
- Compare two means: The null hypothesis for this test is that the means of two samples are equal.
- w = Wilcoxon test statistic
  - Entries for each group ranked (1 = smallest, n1 + n2 = largest)
  - W is dependent on the sum of ranks in a group
  - The Wilcoxon distribution is a discrete, bell-shaped, non-negative distribution
  - W is used to calculate the p-value, which is a probability that measures the evidence against the null hypothesis.

Wilcoxon rank sum test example

- Say we have two samples: sample X has m=3 data points and sample Y has n=4 data points.
- Consider if we combine the samples, order the data points and rank the result.
- Under H0, X & Y have equal means so in the ordered and ranked list of all the data points from the two samples, the data points would be intermingled. There would be no tendency for the sum of the X sample ranks (eg. W) to be higher or lower than Y.
- So, how do we calculate the sampling distribution of W?
- Again, we need to think about what is happening if H0 is true. In this case there is no tendency for the X ranks to be higher or lower than Y so there is an equal chance of a set of 3 ranks in our example say (1,4,5), (3,5,6) or (5,6,7).
- There are 7 choose 3 possible rank triples for X in our example, so we simply calculate W (e.g. the sum of the triple) for each of these and make a histogram.

Probability distribution of W (m=3, n=4) when H0 is true

w	6	7	8	9	10	11	12	13	14	15	16	17	18
P(W=w)	1/35	1/35	2/35	3/35	4/35	4/35	5/35	4/35	4/35	3/35	2/35	1/35	1/35



Discrepancy between R and example

- R uses a slightly different calculation due to Mann and Whitney
- Their test statistic, sometimes called U, is a linear function of the original rank sum statistic, usually called W:

$$U = W - \frac{n_2(n_2 + 1)}{2}$$

- where n2 is the number of observations in the other group whose ranks were not summed.
- This is in fact how the `wilcox.test` and `dwilcox` function calculates the test statistic, though it labels it W instead of U
- Note from the R documentation:

  "The literature is not unanimous about the definitions of the Wilcoxon rank sum and Mann-Whitney tests. The two most common definitions correspond to the sum of the ranks of the first sample with the minimum value subtracted or not: **R**subtracts and S-PLUS does not, giving a value which is larger by *m(m+1)/2* for a first sample of size *m*. (It seems Wilcoxon's original paper used the unadjusted sum of the ranks but subsequent tables subtracted the minimum.)
  R's value can also be computed as the number of all pairs (x[i], y[j]) for which y[j] is not greater than x[i], the most common definition of the Mann-Whitney test."

Additional Resources.

- The Analysis of biological data by Whitlock and Schlutler
- Probability and Statistics for Engineering and the Sciences, Devore 1995.
- An introductory R tutorial made by the Brain Circuits Research Cluster (link here)

Bootstrapping and Permutation Tests

UBC Databinge Data Champions Team

July 13th, 2022

Pulling Ourselves Up By Our Own Bootstraps.

- A phrase that refers to the completely absurd.
- Despite its namesake, many empirical tests have found that bootstrapping works incredibly well.

What is bootstrapping?

- A process where you estimate the sampling distribution of a statistic by resampling from your data with replacement.
  1. You sample a population and get data
  2. You randomly choose points from your data (resample) and allow yourself to take the same points multiple times (replacement).
  3. You repeat step two many, many times and get thousands of resamples.
  4. You calculate a sample statistic from each resample (eg the mean) and then build a distribution of your sample statistics.
- This distribution is called the bootstrap distribution, and it is used as an estimate for the sampling distribution.

The Idea Behind Bootstrapping

"The original sample represents the population from which it was drawn. So, resamples from this sample represent what we would get if we took many samples from the population. The bootstrap distribution of a statistic, based on many resamples, represents the sampling distribution of the statistic, based on many samples"

Companion Chapter 18: Bootstrap Methods And Permutation Tests for the Practice of Business Statistics. David S. Moore, George P. McCabe, William M. Duckworth II, Stanley L. Sclove

Why Use Bootstrapping?

- A great way to explore your data.
- Does not require assumption of normality.
- Less sensitive to sample sizes (but not insensitive).
- Can be generalized to explore statistics other than the mean.
- Determine SEs and CIs.
- If you are interested in the above check the past workshops!
- **Perform permutation tests.**

Assumptions of Permutation Test.

- Random sampling from both populations (of course).
- The populations have identical distributions under the null hypothesis
  - Same mean
  - Same spread (standard deviation)
  - Same shape
- NO assumption of normality.
- Show the alpaca website! https://www.jwilber.me/permutationtest/

Let's Give It A Try.

Additional Resources.

- Companion Chapter 18: Bootstrap Methods And Permutation Tests for the Practice of Business Statistics.

Power Calculations

UBC Databinge Data Champions Team

July 13th, 2022

Power and sample size calculations

- Used to estimate how many samples will be needed for a study or how many subjects are needed to answer the research question.
- Important at the outset of a project when you are setting up your study and often journals will ask how your group sizes were chosen when you are in the final stages of publication.
- Usually these rely on prior experience and previously collected data as you need an estimate for the variance you expect in your samples.
- Think of these calculations as numerical experiments that can help guide experiment design
- R has a built in package called power to assist with this so we will show a few examples.

ANOVA

UBC Databinge Data Champions Team

July 13th, 2022

More than 2 treatment groups.

- Two sample t-tests are great, but they can only compare between two groups by design.
- What if we want to do 3 or more groups?
- Example: You have three groups of mice. One group receives a half dose of a drug, one group receives a full dose, and the last group is a control.

Control

Half Dose

Full Dose

Note: We cannot just do multiple t-tests, more on this later.

ANOVA.

- ANOVA = analysis of variance.
- In ANOVA, your null hypothesis is that all of the treatment groups have the same mean value for our variable of interest.

$$\mu_1=\mu_2=\mu_3$$

- Your alternative hypothesis is that at least one of the treatment groups has a significantly different mean

$$\mu_1=\mu_2\neq\mu_3 \text{ or } \mu_1\neq\mu_2=\mu_3 \text{ or } \mu_1\neq\mu_2\neq\mu_3$$

NOTE: ANOVA does not show us which groups are different from each other, it just tells us that at least one group is significantly different from the rest.

The Idea of Anova.

- If all of the groups are the same, then how much variation do we expect to see between the groups due to sample variation.
- If there are no differences between groups then taking a random sample from each group is the same as taking one big sample from one group.
- We treat all the datapoints like they were from one big sample and then analyze the variation within that combined sample. If there is much more variation than we would expect under the null hypothesis, then we conclude that at least one of the groups is different which is the source of the extra variation.
- How do we do this practically?

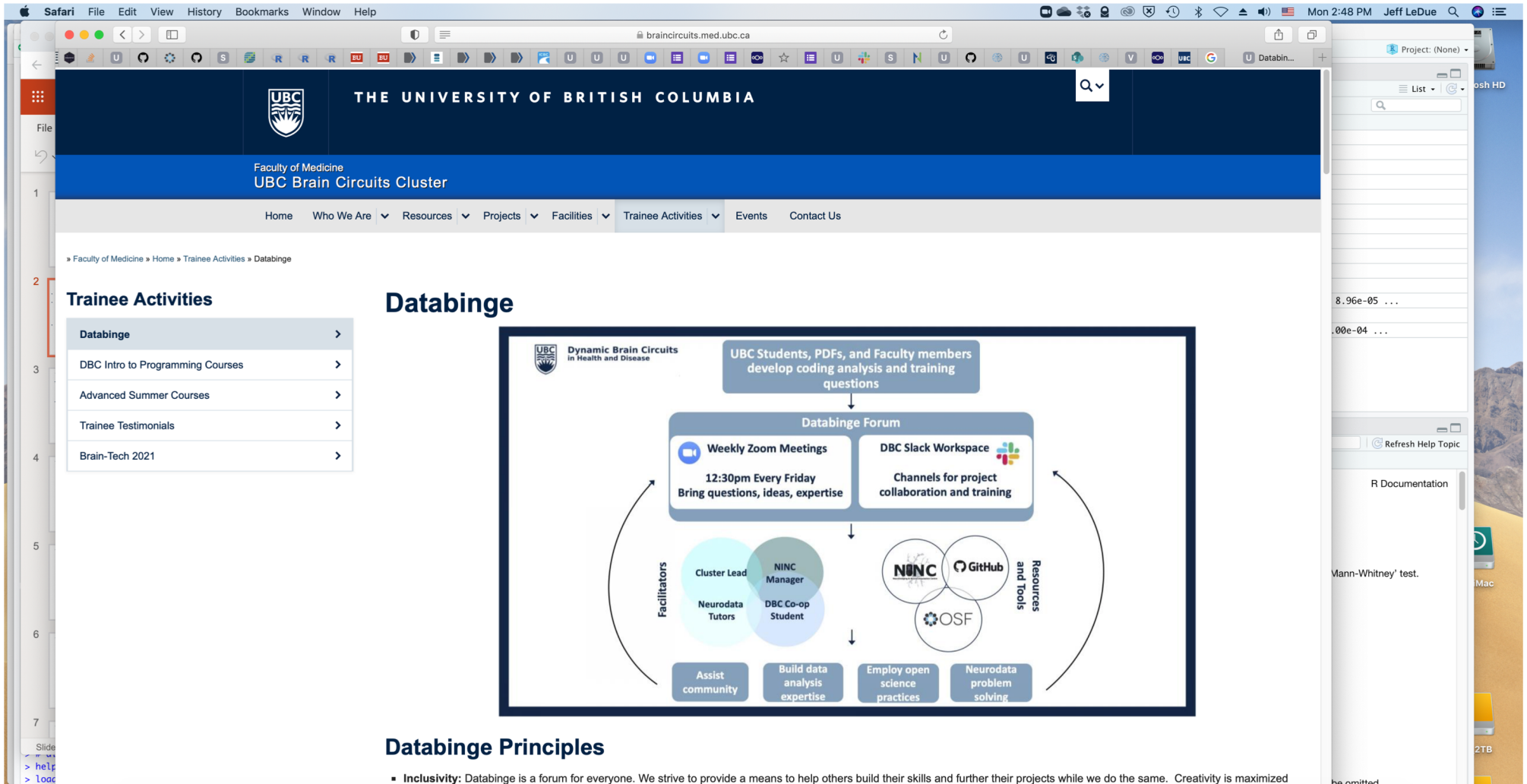
How to do ANOVA.

- There are at least 8 different values that need to be calculated to perform ANOVA.
- The calculations are not that difficult, but they are time consuming, and it is easy to get lost among the different steps.
- Practically speaking, ANOVA is always performed using coding software like R.
- Let's jump into R and give it a try.

To consider for Part 3:

- Theory of ANOVA, F-statistic, F distribution
- The ANOVA table
- Assumptions of ANOVA
- Why can we not just do multiple tests?
- Family wise error rate, False discovery rate
- Multiple comparisons.

Where can I get more support?



Thanks all!