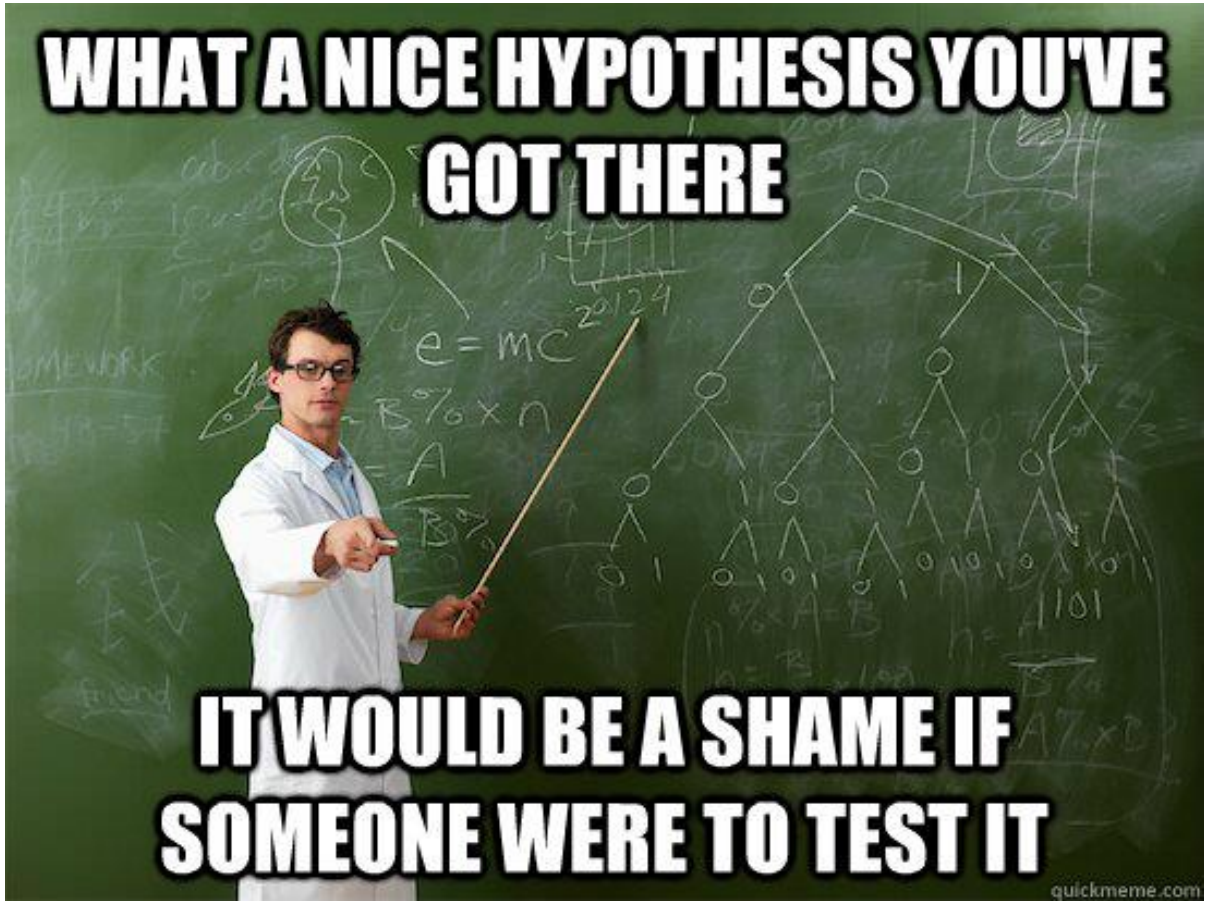
Randomization tests continued



Overview

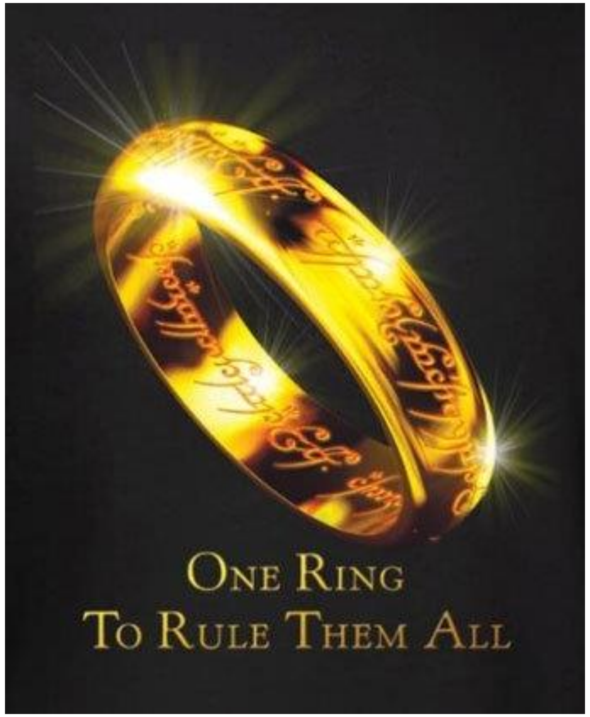
Quick review of hypothesis tests for a single proportion

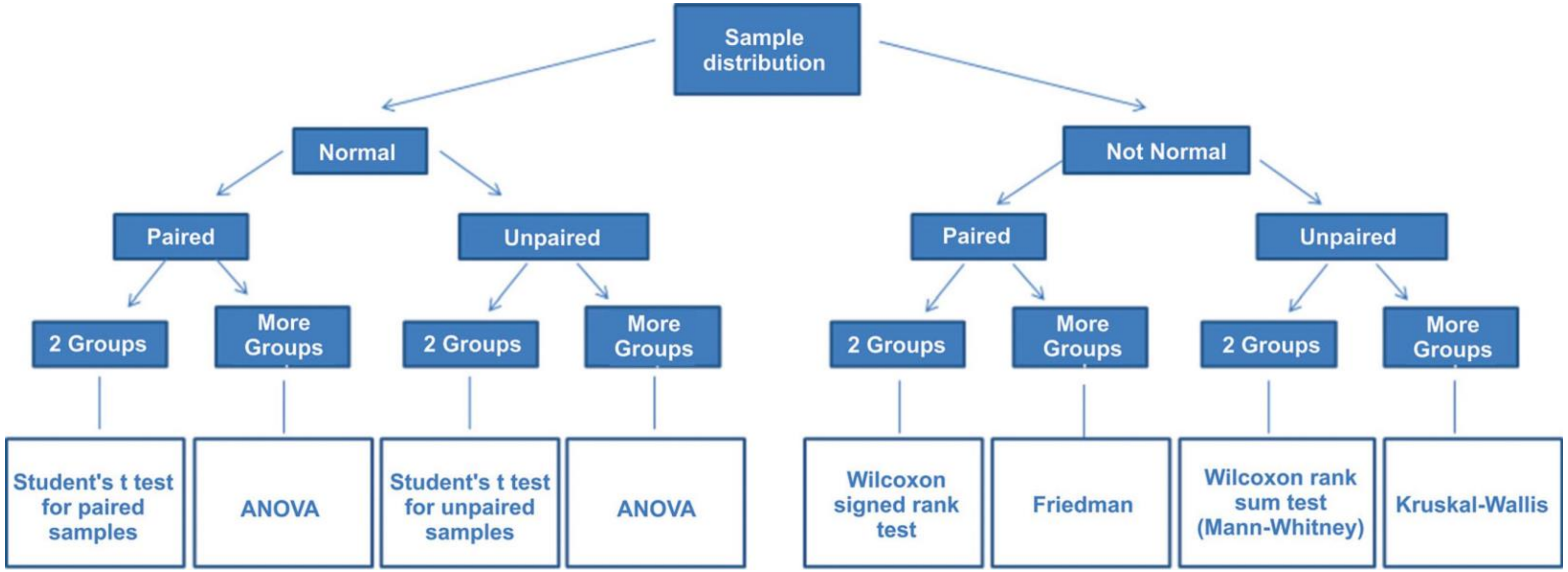
Randomization tests for two means

Randomization tests for more than two means

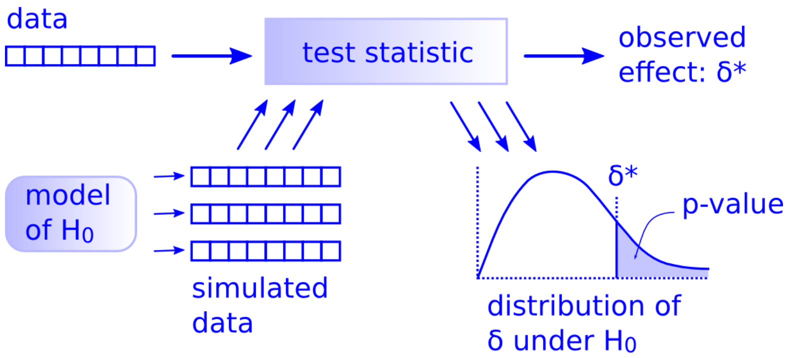
If there is time: theories of hypothesis testing

The big picture: There is only one hypothesis test!

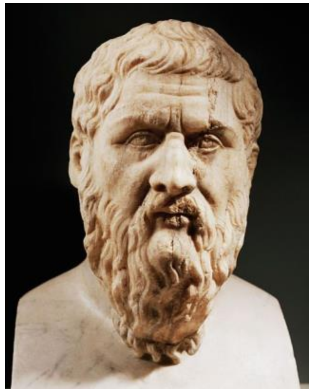




Just need to follow 5 steps!



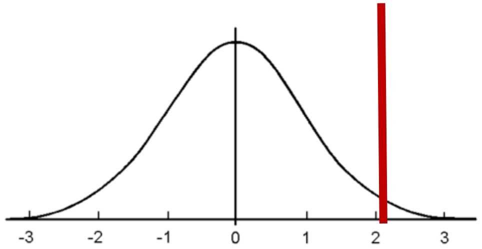
Five steps of hypothesis testing



1. State H_0 and H_A
    - Assume Gorgias (H_0) was right
    - α = .05 of the time he will be right, but we will say he is wrong

2. Calculate the actual observed statistic

3. Create a distribution of what statistics would look like if Gorgias is right
    - Create the **null distribution** (that is consistent with H_0)



4. Get the probability we would get a statistic more than the observed statistic from the null distribution
    - p-value

5. Make a judgement
    - Assess whether the results are statistically significant

Review: hypothesis test for a single proportion

Joy Milne claimed to have the ability to smell whether someone had Parkinson's disease

To test this claim researchers gave Joy 6 shirts that had been worn by people who had Parkinson's disease and 6 shirts by people who did not.

Joy identified 11 out of the 12 shirts correctly.

Step 1: state the null and alternative hypotheses

- H_0: $\pi = 0.5$
- H_A: $\pi > 0.5$ ← H_0 and H_A need to be mutually exclusive

Review: hypothesis test for a single proportion

We can run a hypothesis test for a single proportion in R using:

```
obs_stat <- 11/12          #  Step 2: calculate the observed statistic

flip_sims_prop <- rbinom(10000, 12, .5)/12     # Step 3: create null distribution

p_value <- sum(flip_sims_prop >= obs_stat)/length(flip_sims)     # Step 4: p-value
```

p-value is   0.0029

Step 5: Should we reject H_0?

Do you really believe Joy can smell Parkinson's disease?

TREATMENTS

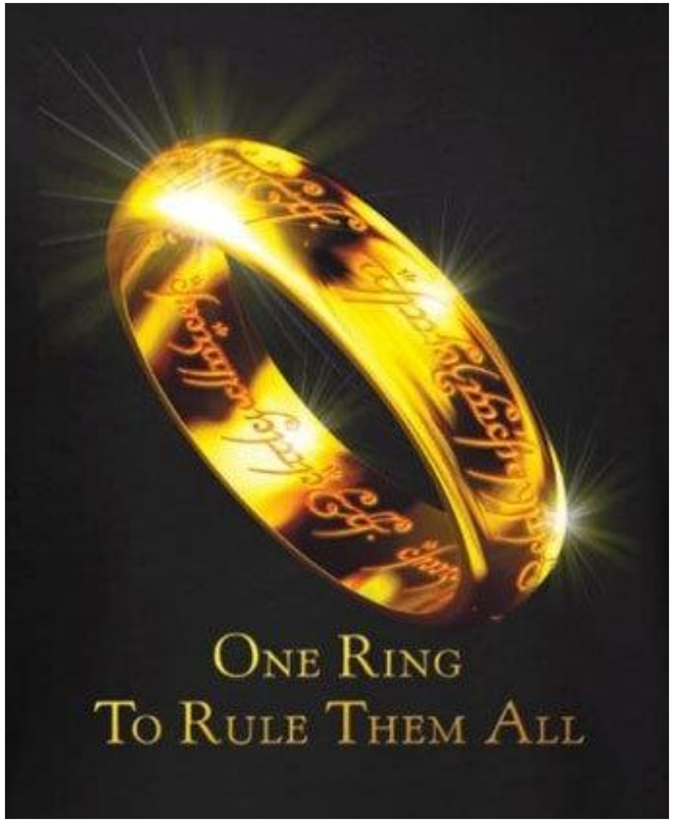
Her Incredible Sense Of Smell Is Helping Scientists Find New Ways To Diagnose Disease

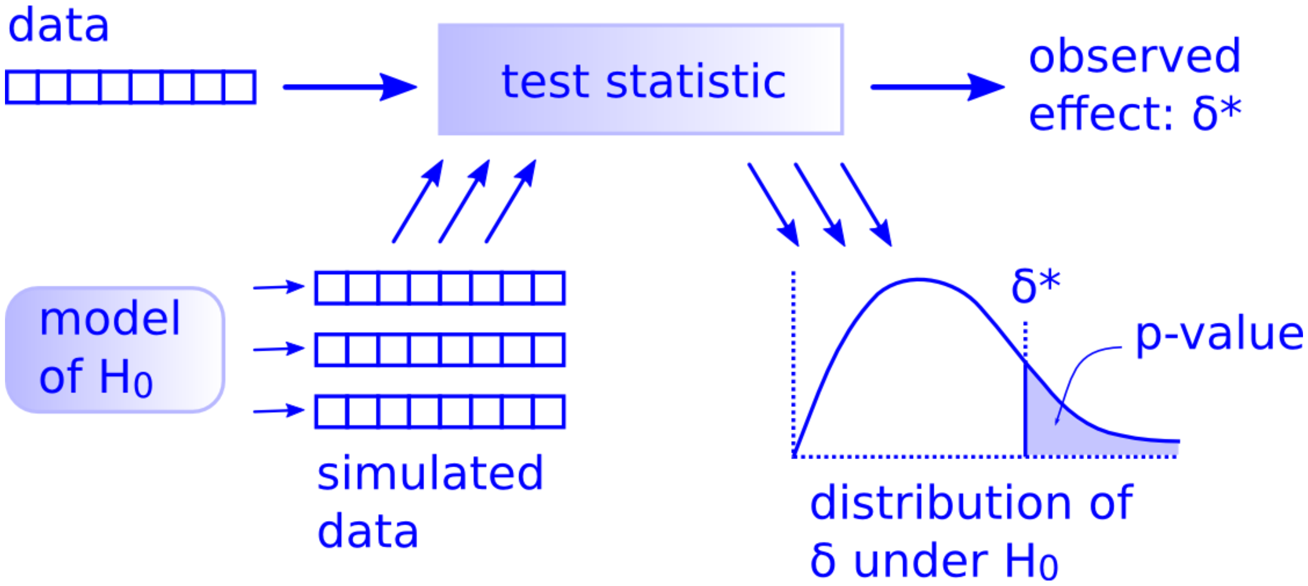
March 23, 2020 · 4:45 PM ET

Questions?

Hypothesis tests comparing 2 means

The big picture: There is only one hypothesis test!





Just need to follow 5 steps!

Hypothesis tests for comparing two means

Question: Is this pill effective?

Testing whether a pill is effective (on average)

How would we design a study?

What would the cases and variables be?

What would the parameter and statistic of interest be?

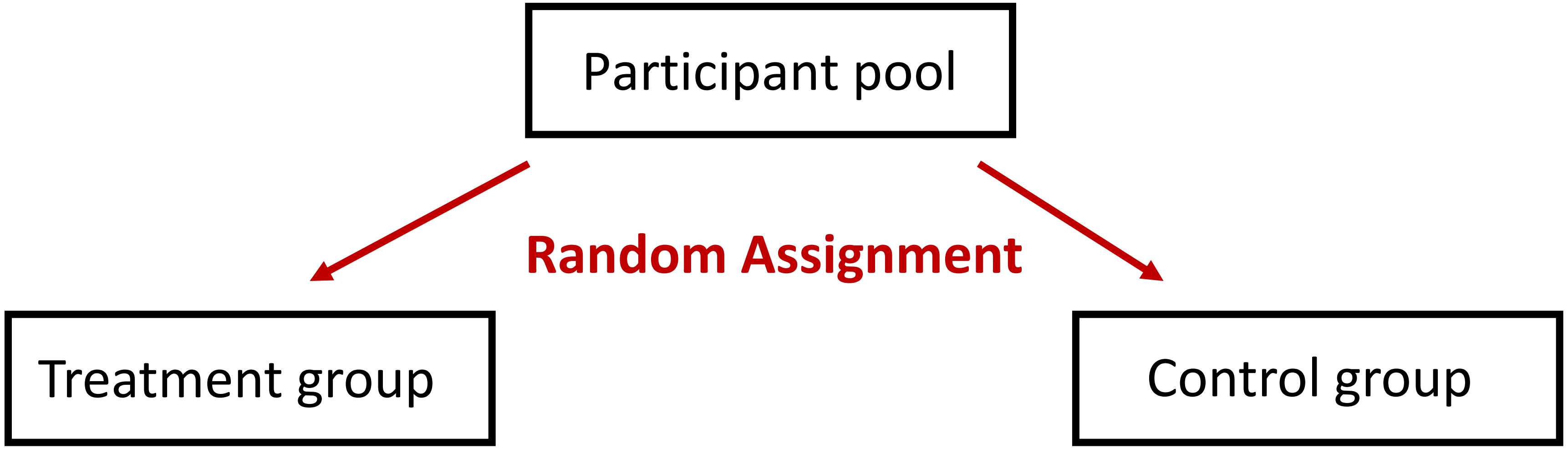
What are the null and alternative hypotheses?

- Assume we are looking for differences in means between the groups

Experimental design

Take a group of participant and ***randomly assign***:

- Half to a *treatment group* where they get the pill
- Half in a *control group* where they get a fake pill (placebo)
- See if there is more improvement in the treatment group compared to the control group



Observational and experimental studies

An **observational study** is a study in which the researcher does not actively control the value of any variable but simply observes the values as they naturally exist.

An **experiment** is a study in which the researcher actively controls one or more of the explanatory variables

- **Random assignment** is where experimental units are randomly assigned to treatment and control groups which allows one to answer questions about **causation**!

Question: Which data are from observational studies?

- Most drug studies
- OkCupid data
- Joy Smelling Parkinson's

Hypothesis tests for differences in two group means

1. State the null and alternative hypothesis
   - H_0: $\mu_{Treatment} = \mu_{Control}$ or $\mu_{Treatment} - \mu_{Control} = 0$
   - H_A: $\mu_{Treatment} > \mu_{Control}$ or $\mu_{Treatment} - \mu_{Control} > 0$

2. Calculate statistic of interest
   - $\bar{x}_{Effect} = \bar{x}_{Treatment} - \bar{x}_{Control}$

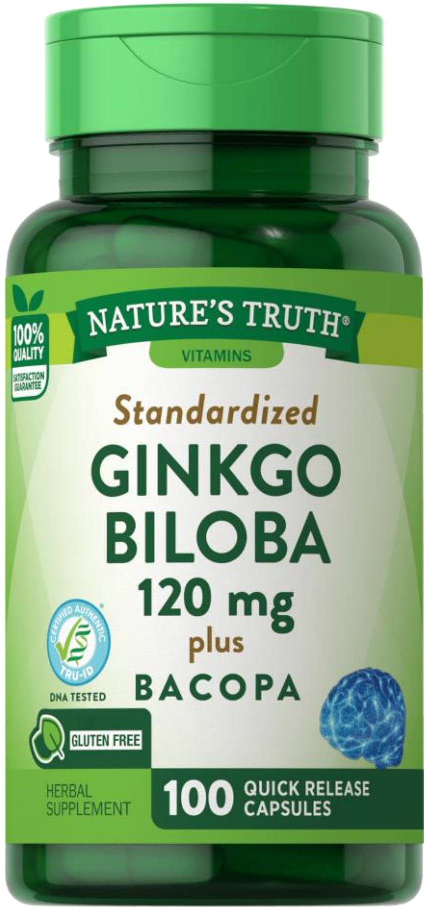
Example: Does Gingko improve memory?

A double-blind randomized controlled experiment by Solomon et al (2002) investigated whether taking a Ginkgo supplement could improve memory

- A treatment group of n = 104 participants took a Gingko supplement 3 times per day for 6 weeks
- A control group of n = 99 participants took a placebo 3 times per day for 6 weeks

Standardized neuropsychological tests of learning and memory, attention and concentration were measured at the end of the six week period.

Question: Was there a difference in the mean cognitive score between the treatment and control groups?



1. State the null and alternative hypothesis

In words:

- Null hypothesis: The average memory score will be the same for participants who took Gingko and the placebo
- Alternative hypothesis: The average memory score will be different for the two groups.

In symbols:

- H_0: $\mu_{Treatment} = \mu_{Control}$ or $\mu_{Treatment} - \mu_{Control} = 0$
- H_A: $\mu_{Treatment} \neq \mu_{Control}$ or $\mu_{Treatment} - \mu_{Control} \neq 0$

2. Visual the data can calculate the observed statistic

How could we visualize the data?

- We will try this in R soon…

What could we use for the observed statistic?

3. Create the null distribution!

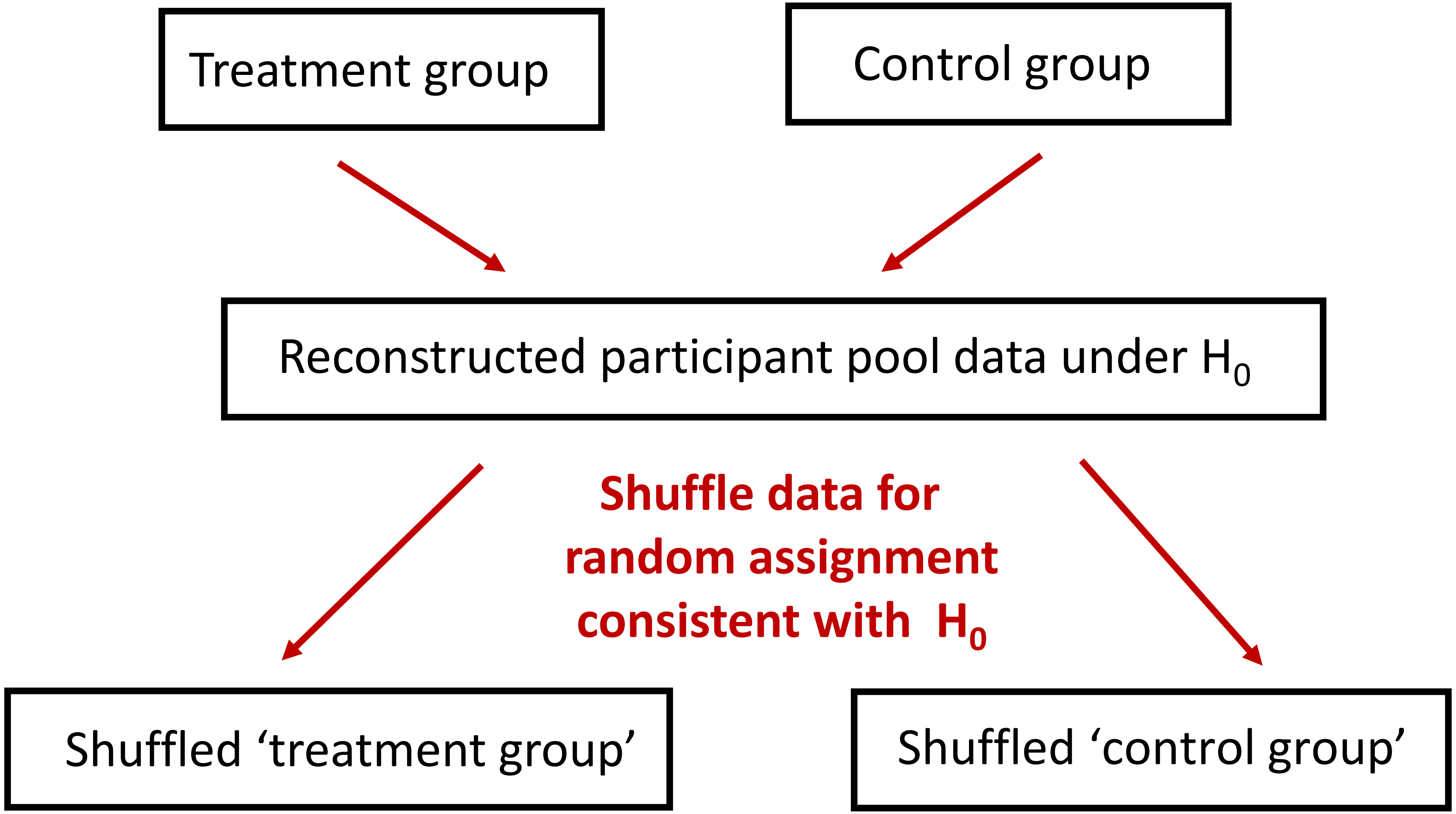
How could we create the null distribution?

Need to generate data consistent with H_0: $\mu_{Treatment} - \mu_{Control} = 0$

- i.e., we need fake $\overline{x}_{Effect}$ that are consistent with H_0

Any ideas how we could do this?

3. Create the null distribution!



One null distribution statistic: $\bar{x}_{Shuff\_Treatment} - \bar{x}_{Shuff\_control}$

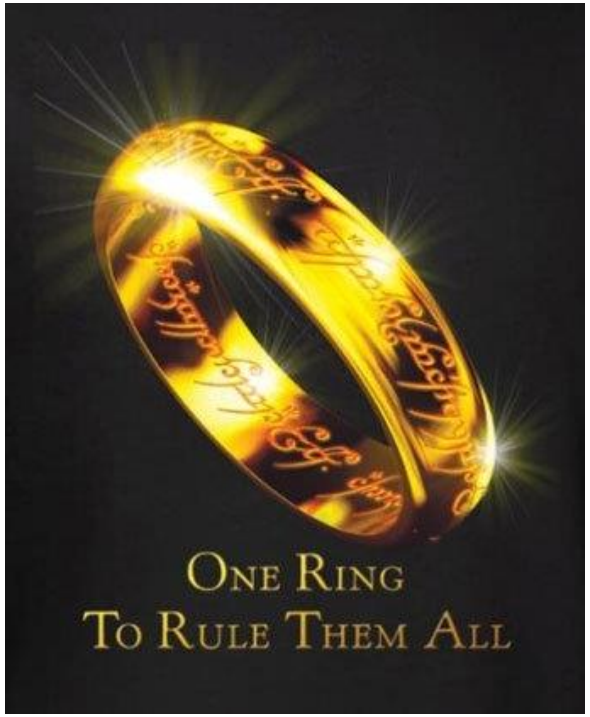
3. Create a null distribution

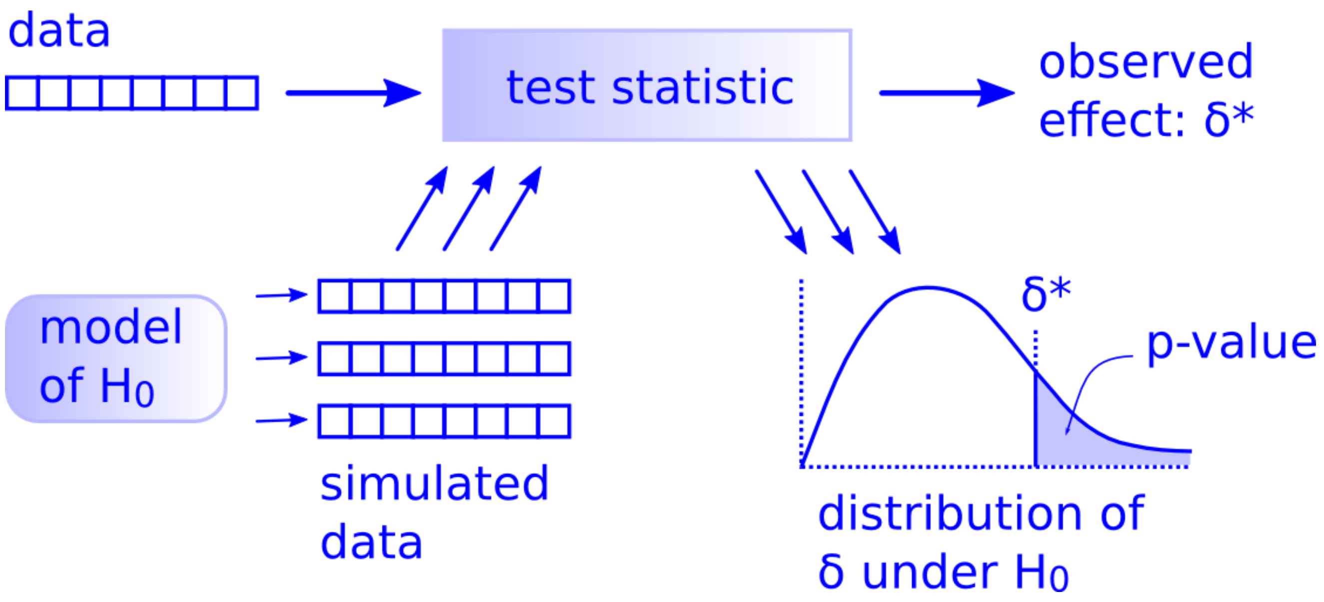
1. Combine data from both groups
2. Shuffle data
3. Randomly select 104 points to be the ‘null’ treatment group
4. Take the remaining 99 points to the ‘null’ control group
5. Compute the statistic of interest on these ‘null’ groups
6. Repeat 10,000 times to get a null distribution

Let’s try the rest of the hypothesis test in R…

Hypothesis test for comparing more than two means

The big picture: There is only one hypothesis test!

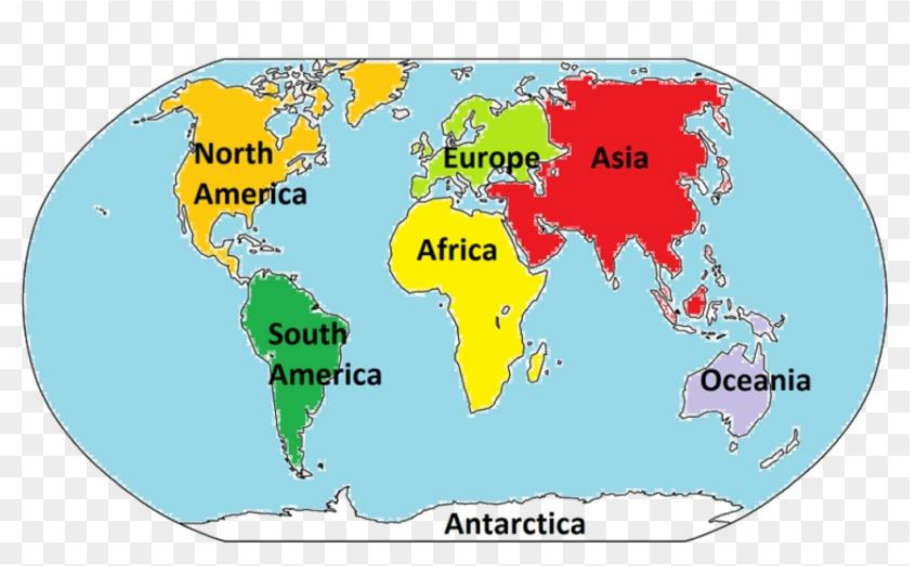




Just need to follow 5 steps!

Comparing more than two means

Let's examine the beer consumption in different continents!



Analysis inspired by:

- [Minitab blog article]
- [Five thirty eight analysis]

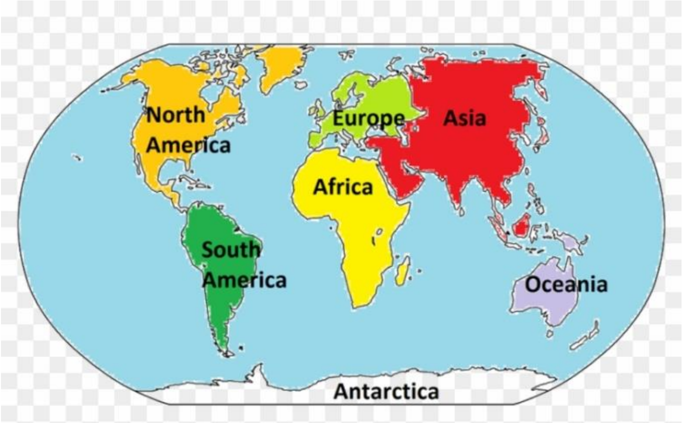
Question: Does the average beer consumption in countries different depending on the continent?

1. State the null and alternative hypotheses!

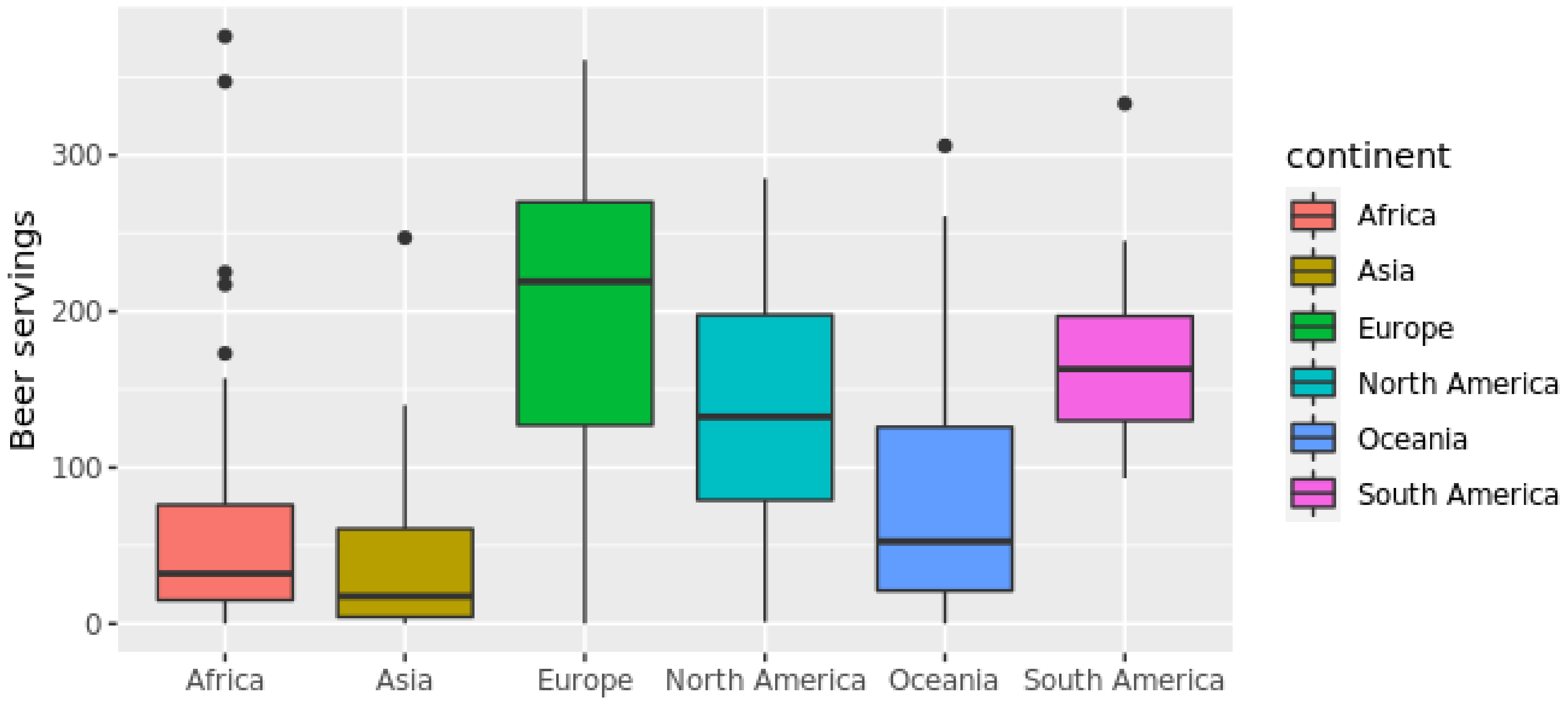
H_0: $\mu_{Asia} = \mu_{Europe} = \mu_{Africa} = \mu_{North\text{-}America} = \mu_{South\text{-}America} = \mu_{Oceania}$

H_A: $\mu_i \neq \mu_j$ for at least one pair of fields of continents

What should we do next?



Plot of the beer consumption in different continents



Thoughts on the statistic of interest?

Comparing multiple means

There are many possible statistics we could use. A few choices are:

1. Group range statistic:

   max $\bar{x}$ - min $\bar{x}$

2. Mean absolute difference (MAD):

   $(|\bar{x}_{Africa} - \bar{x}_{Asia}| + |\bar{x}_{Africa} - \bar{x}_{Europe}| + \ldots + |\bar{x}_{Oceania} - \bar{x}_{South\text{-}America}|)/15$

3. F statistic:

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

Using the MAD statistic

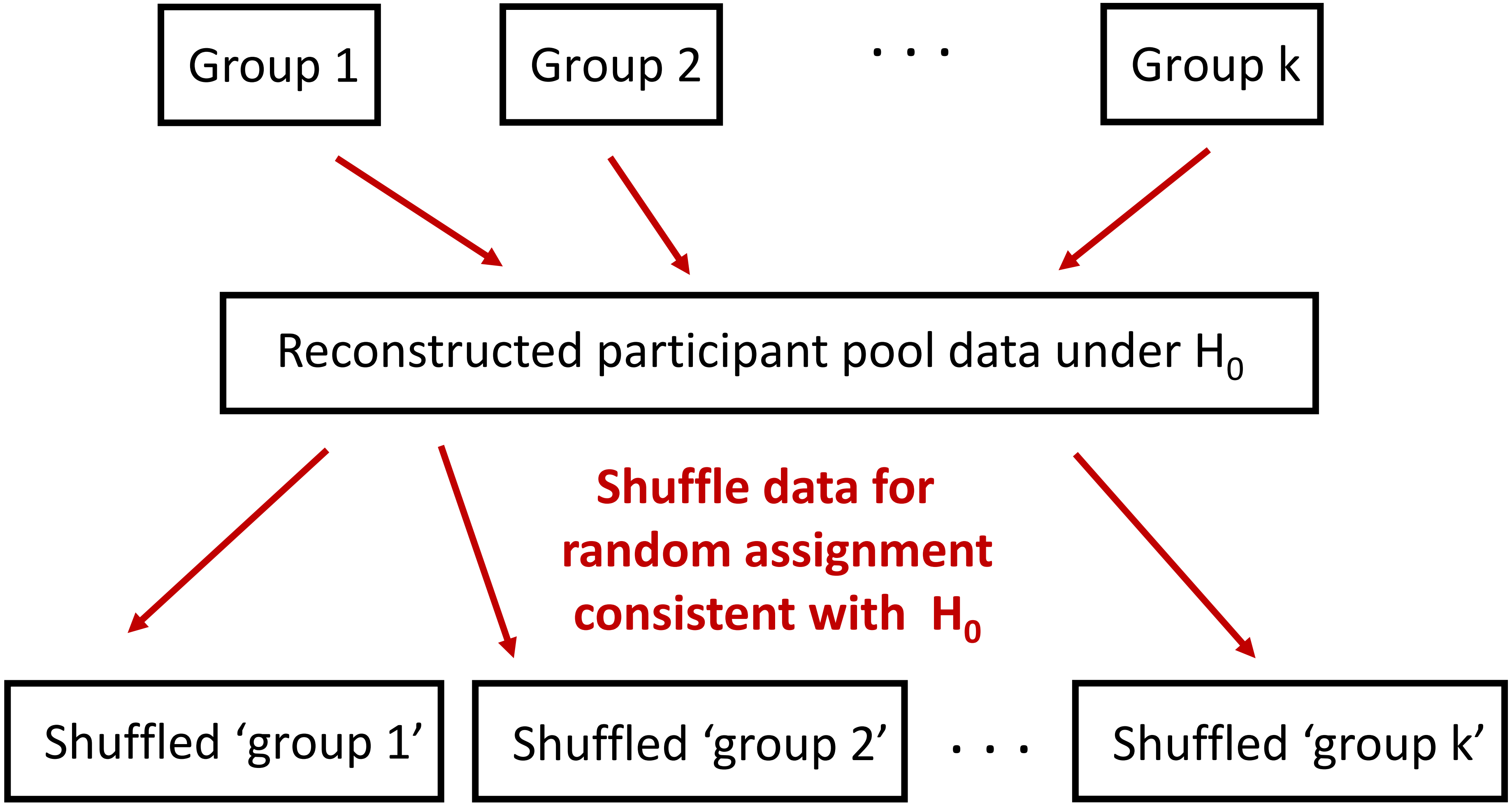
<u>Mean absolute difference (MAD):</u>

$$(|\bar{x}_{Africa} - \bar{x}_{Asia}| + |\bar{x}_{Africa} - \bar{x}_{Europe}| + \ldots + |\bar{x}_{Oceania} - \bar{x}_{South\text{-}America}|)/15$$

Observed statistic value = 78.86

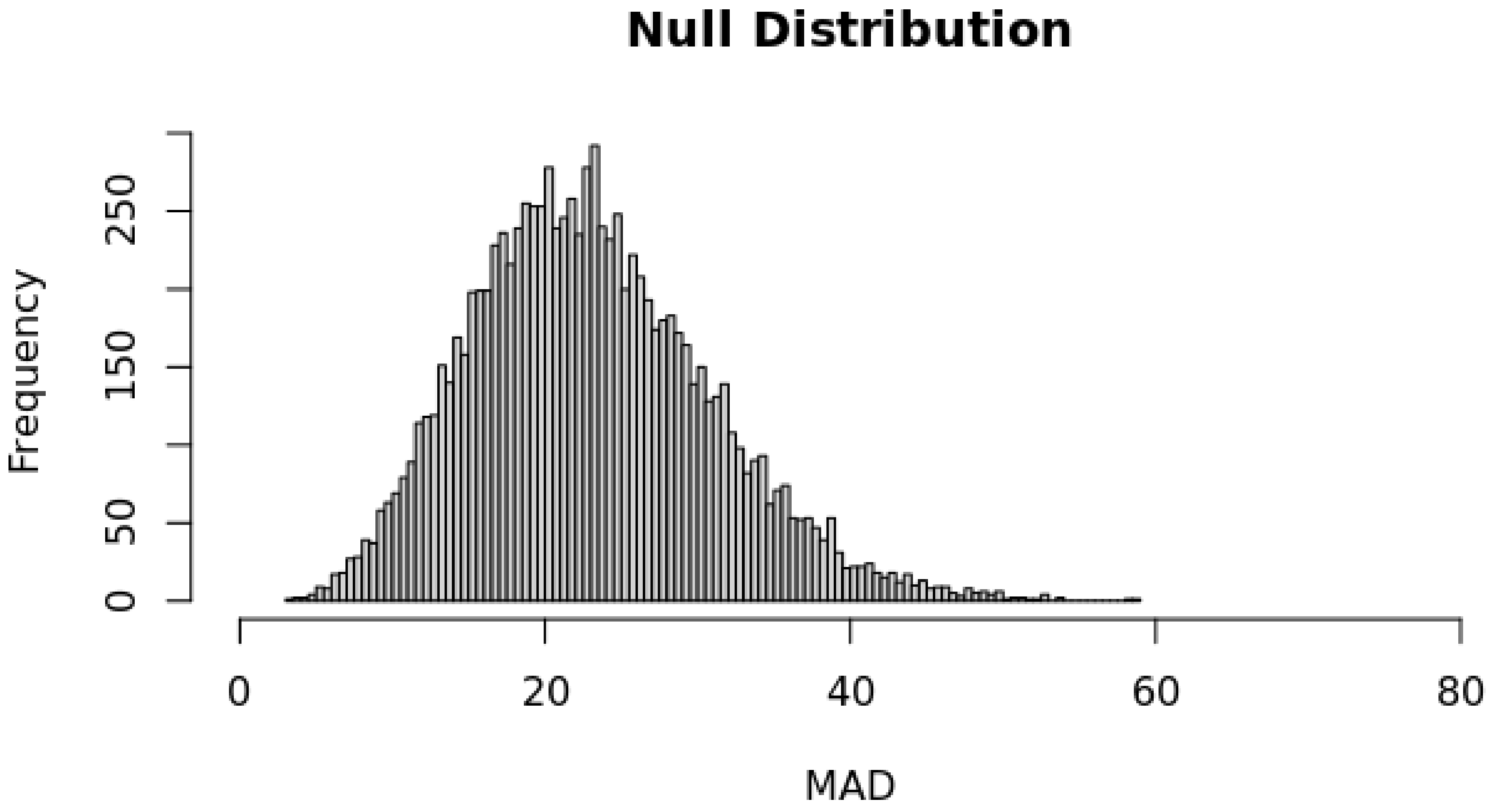
How can we create the null distribution?

3. Create the null distribution!

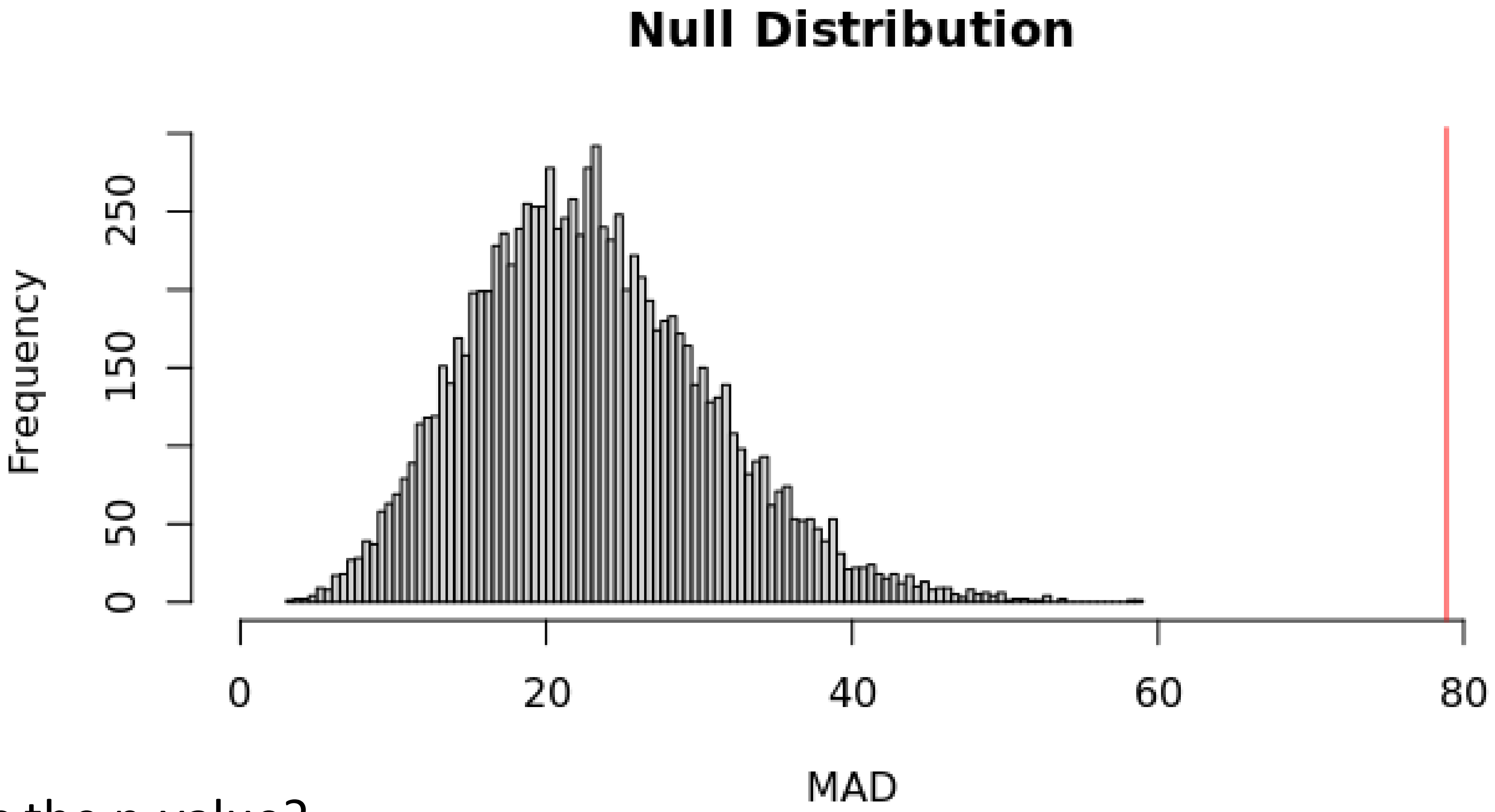


Compute statistics from shuffled groups

3. Create the null distribution!



4. Calculate the p-value

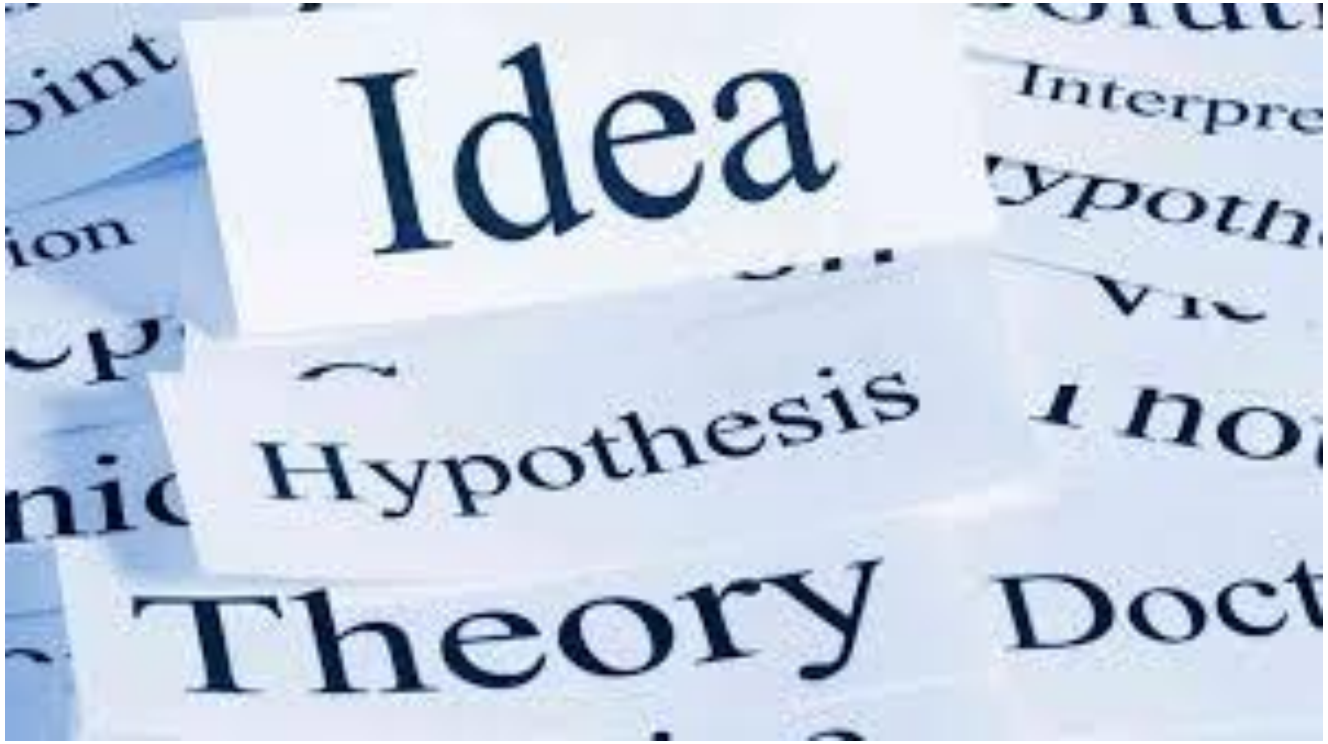


What is the p-value?

Conclusions?

Let’s try it in R…

Theories of hypothesis tests



Two theories of hypothesis testing

Null-hypothesis significance testing (NHST) is a hybrid of two theories:

1. Significance testing of Ronald Fisher
2. Hypothesis testing of Jezy Neyman and Egon Pearson

Fisher (1890-1962)

Neyman (1894-1981)

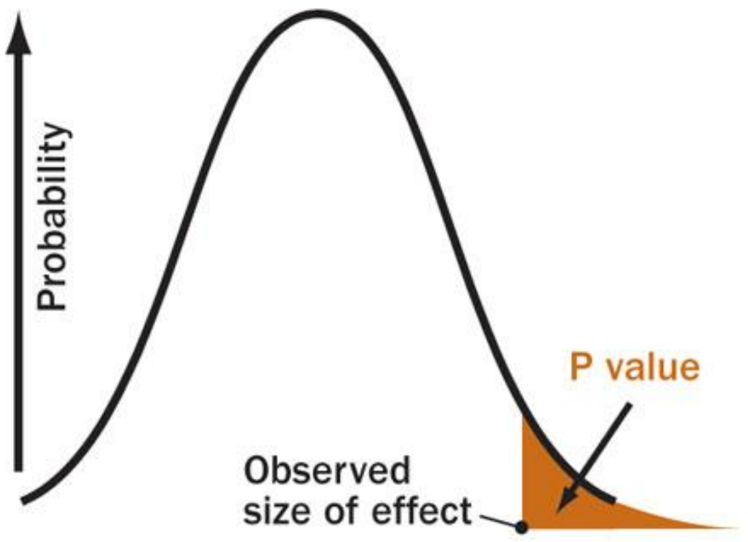
Pearson (1895-1980)

Ronald Fisher’s significance testing

Views the p-value as strength of evidence against the null hypothesis

- p-values part of an on-going scientific process:

  They tell the experimenter “what results to ignore”



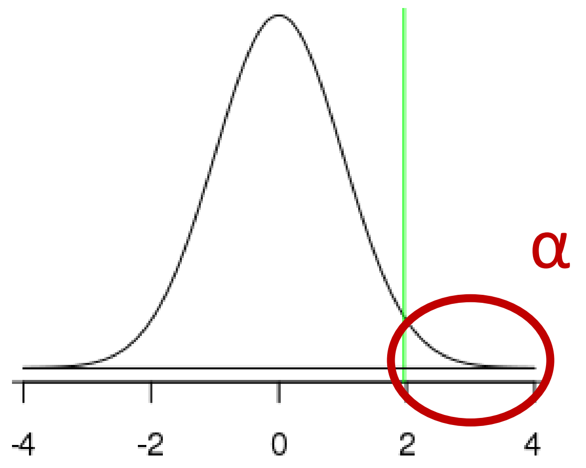
Neyman-Pearson null hypothesis testing

Makes ***a formal decision*** in statistical tests

Reject H_0: if the observed sample statistic is beyond a fixed value

- i.e., reject H_0 if the p-value is less than some predetermined **significance level** α

Null distribution



Do not reject H_0: if the observed sample statistic is not beyond a fixed value. This means the test is inconclusive.

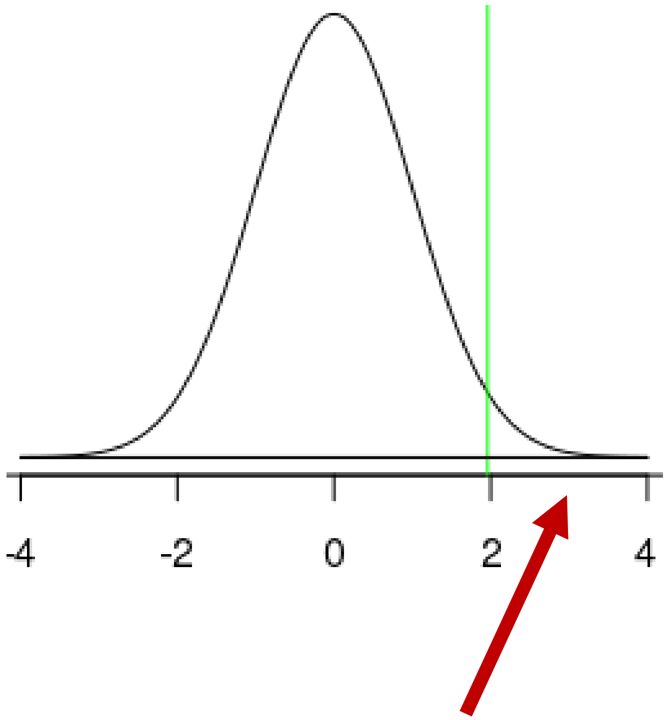
Neyman-Pearson frequentist logic

Type I error: incorrectly rejecting the null hypothesis when it is true

If Neyman-Pearson null hypothesis testing paradigm was followed perfectly, and we were in a world where the null hypothesis was always true, then only ~5% of the time would we falsely report an effect (for $\alpha = 0.05$)

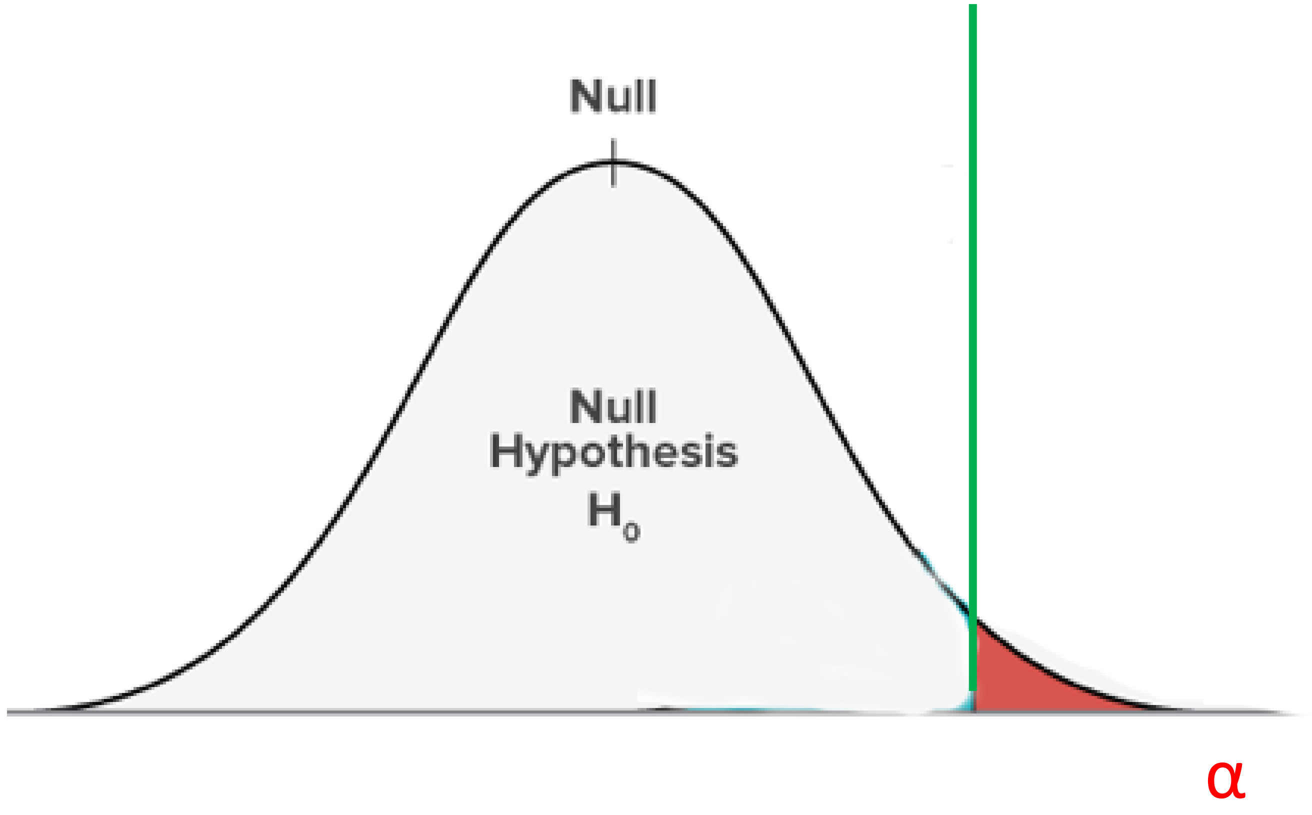
- i.e., we would only make type I errors 5% of the time

Null distribution

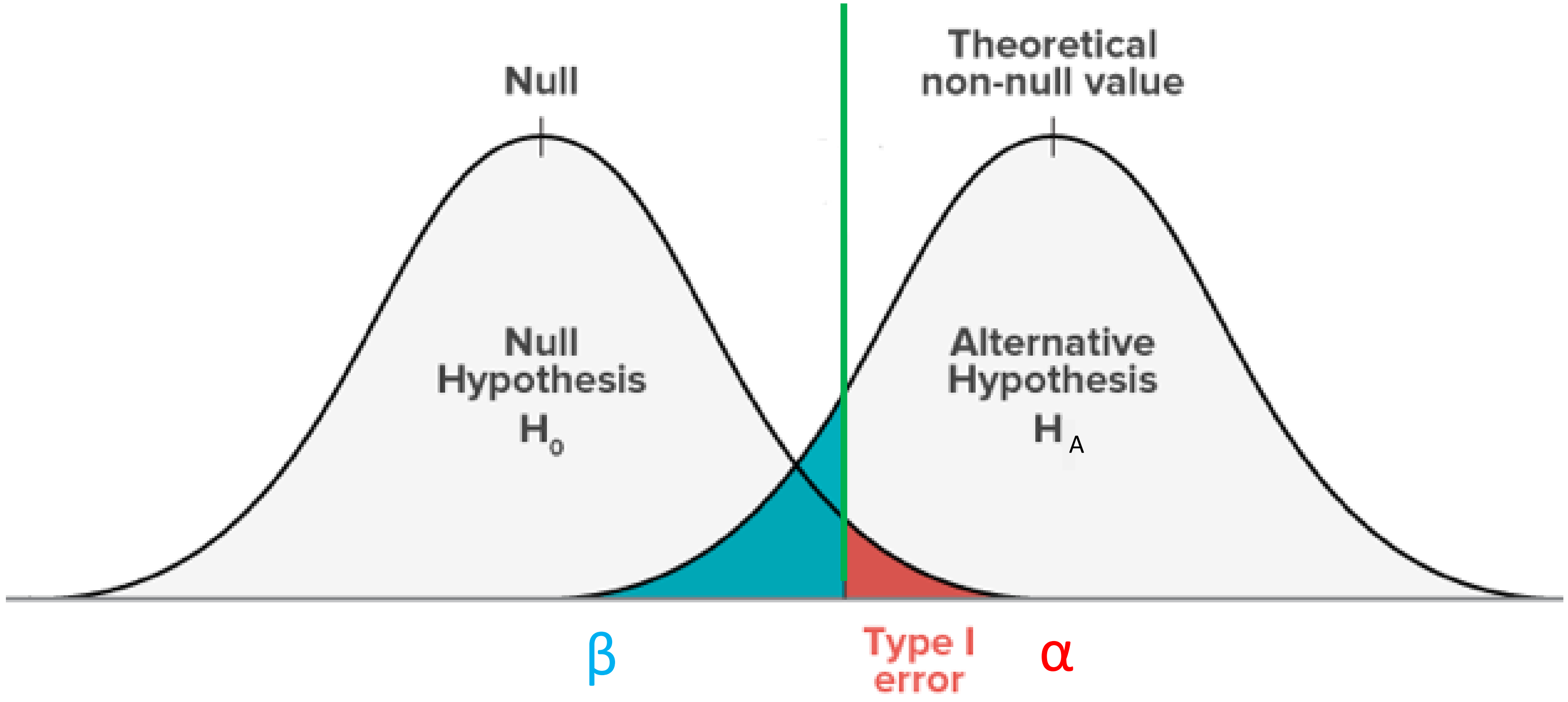


The null distribution is true but statistic landed here

Neyman-Pearson Frequentist logic



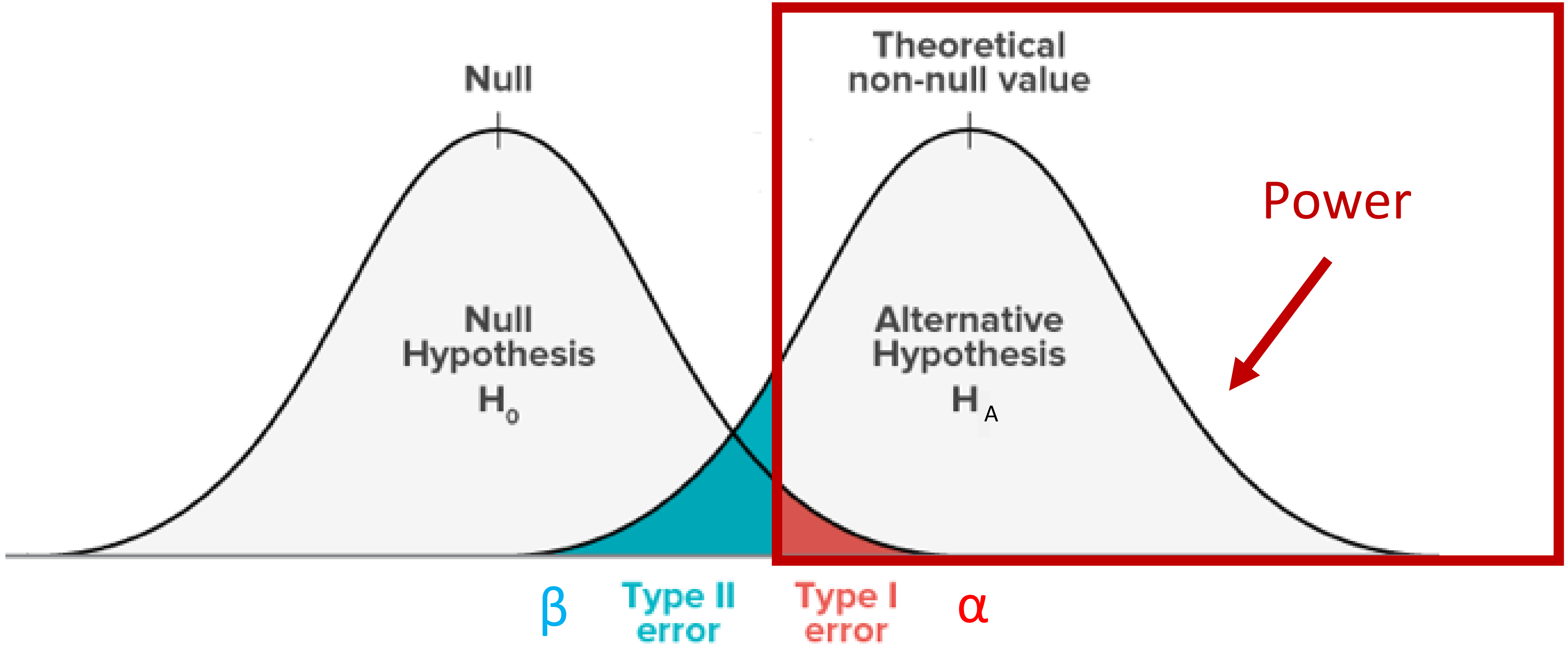
Neyman-Pearson Frequentist logic



Type II error: incorrectly rejecting failing to reject H_0 when it is false

- The rate at which we make type II errors is often denoted with the symbol β

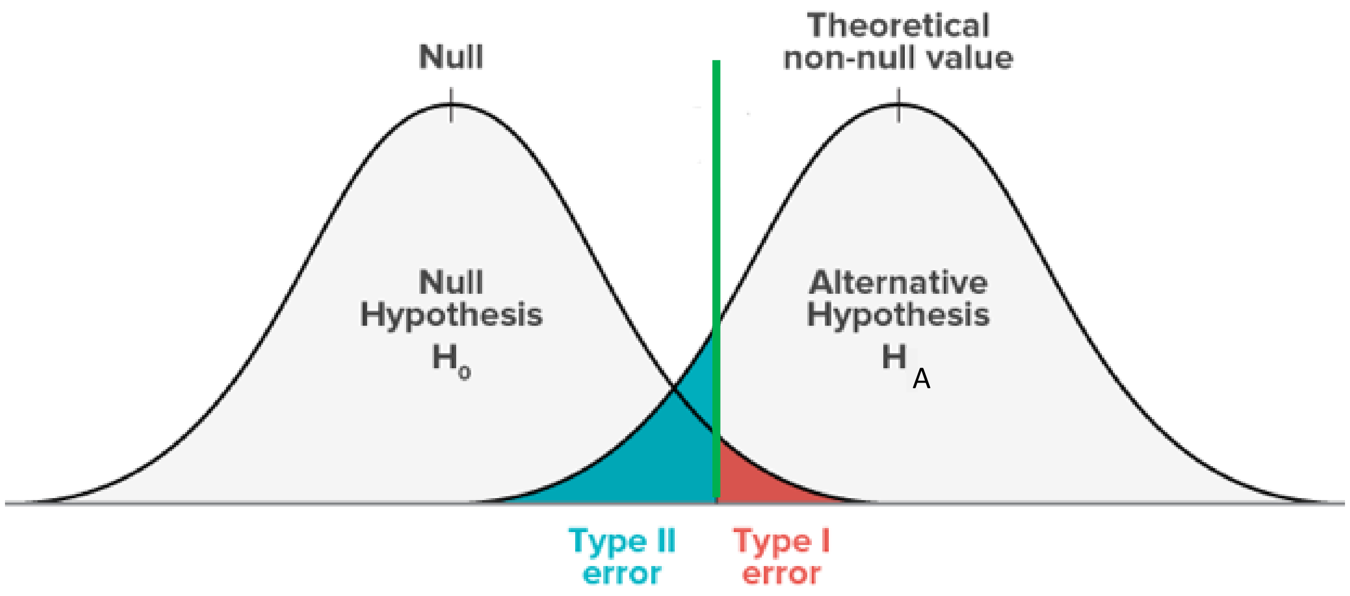
Neyman-Pearson Frequentist logic



The **power** of a test is the probability we reject the H_0 when it is **false**

- 1 - β
- For a fixed α level, it would be best to use the most powerful test

Type I and Type II Errors



Decision

	Reject H_0	**Do not reject H_0**
H_0 is true	Type I error (α) (false positive)	No error

Truth

Problems with the NP hypothesis tests

Problem 1: we are interested in the results of a specific experiment, not whether we are right most of the time

- E.g., 95% of these statements are false:
  - Joy can't smell Parkinson's disease, Lawyers are left-handed at the same rate as the general population, Calcium is not beneficial for your heart, …

Problem 2: Arbitrary thresholds for alpha levels

- P-value = 0.051, we don't reject H_0

Problems with the NP hypothesis tests

Problem 1: we are interested in the results of a specific experiment, not whether we are right most of the time

- E.g., 95% of these statements are false:
  - Joy can't smell Parkinson's disease, Lawyers are left-handed at the same rate as the general population, Calcium is not beneficial for your heart, …

Problem 2: Arbitrary thresholds for alpha levels

- P-value = 0.051, we don't reject H_0?

Problem 3: running many tests can give rise to a high number of type I errors

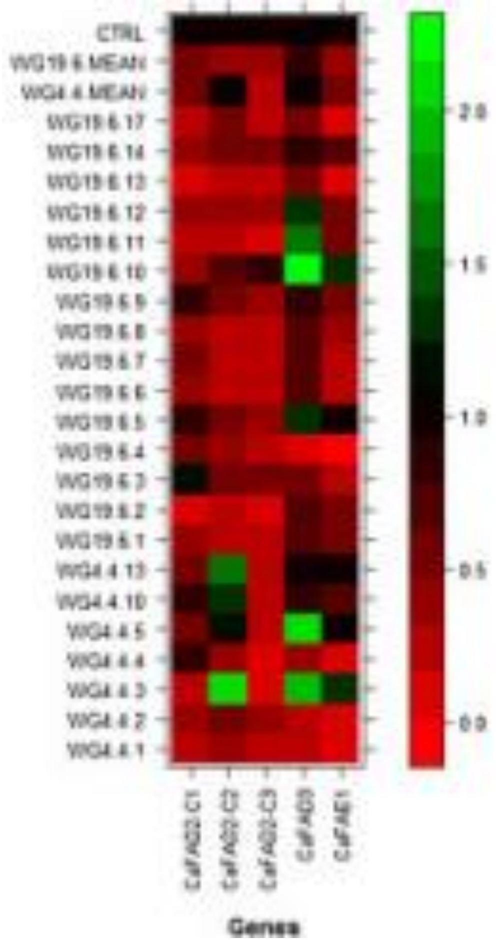
Genes and leukemia example

Scientists collected 7129 gene expression levels from 38 patients to find genetic differences between two types leukemia (L1 and L2)

Suppose there was no genetic differences between the types of leukemia

- H_0: $\mu_{L1} = \mu_{L2}$ is true for all genes

Q: If each gene was tested separately using a significance level of $\alpha = 0.05$, approximately how many type I errors would be expected?



The problem of multiple testing

For $\alpha = 0.05$, ~5% of all published research findings should incorrectly reject the null hypothesis

Publication bias (file drawer effect): Generally positive results are more likely to be published, so if you read the literature, the proportion of incorrect results could be greater than 5%.

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

The Earth Is Round (p < .05)

Jacob Cohen

After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists. This article reviews the problems with this practice, including

sure how to test H_0, chi-square with Yates's (1951) correction or the Fisher exact test, and wonders whether he has enough power. Would you believe it? And would you believe that if he tried to publish this result without a

American Statistical Association's Statement on p-values

Some thoughts…

Better to have hypothesis tests than none at all. Just need to think carefully and use your judgment.

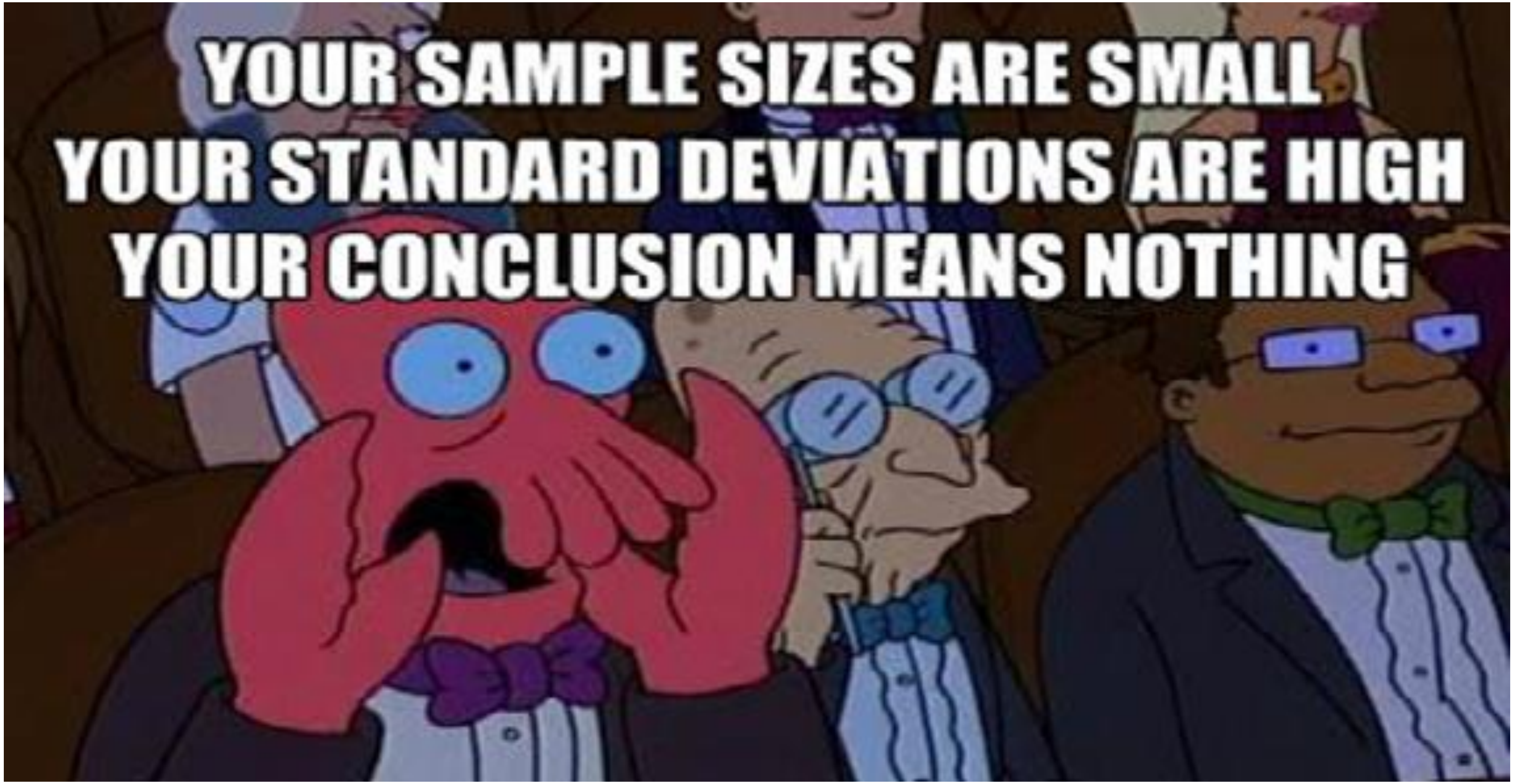
Report effect size in most cases – i.e., confidence intervals

Report the p-values rather than accept/reject H_0

- i.e., report $p = 0.23$ not $p < 0.05$

Replicate findings (perhaps in different contexts) to make sure you get the same results

Be a good/honest scientists and try to get at the Truth!


YOUR SAMPLE SIZES ARE SMALL
YOUR STANDARD DEVIATIONS ARE HIGH
YOUR CONCLUSION MEANS NOTHING

Next week

Parametric hypothesis tests and more…