Parametric hypothesis tests

Overview

Tests for two means

- Randomization tests for two means revisited using a t-statistic
- Parametric tests and t-tests

Theories of hypothesis testing

Announcements

Ed discussions

- Use it!
  - Your class participation grade (3%) based on asking and answering questions
- But do not post full code/answers on Ed Discussion!

Technical issues recording lectures the last two classes

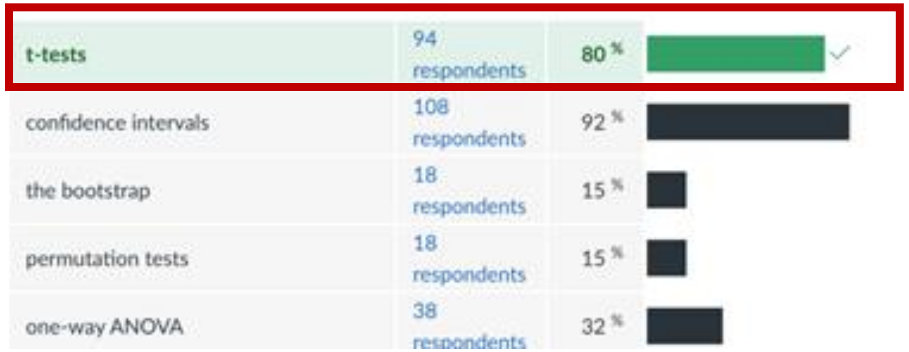
- I posted lecture material from last year and audio recordings
- Hopefully will be fixed this class!

Next programming review will be from 4-5pm next Monday

Where we are in the plan for the semester

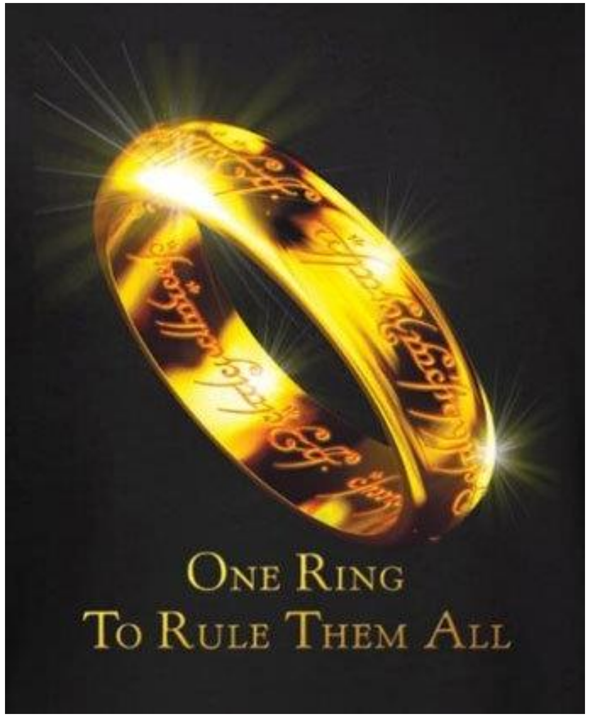
			Analysis	R
1	Sep 2	Course overview, introduction to R, descriptive statistics		base R
2	Sep 7-9	Review of central statistical concepts and exploratory analysis using R	resampling methods	
3	Sep 14-16	Confidence Intervals and the bootstrap		
4	Sep 21-23	Review of hypothesis tests and permutation tests in R		
5	Sep 28-30	Parametric hypothesis tests and theories of hypothesis testing		

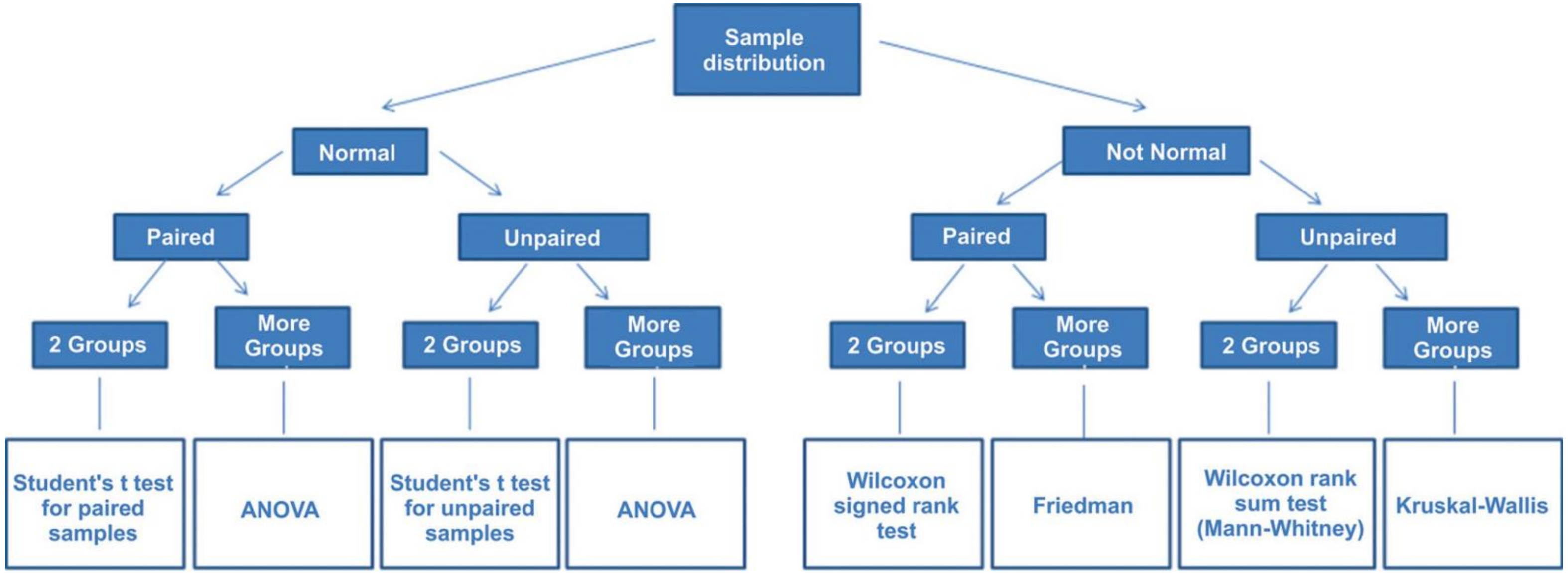
YOU ARE HERE



t-tests	94 respondents	80 %	✓
confidence intervals	108 respondents	92 %	
the bootstrap	18 respondents	15 %	
permutation tests	18 respondents	15 %	
one-way ANOVA	38 respondents	32 %	

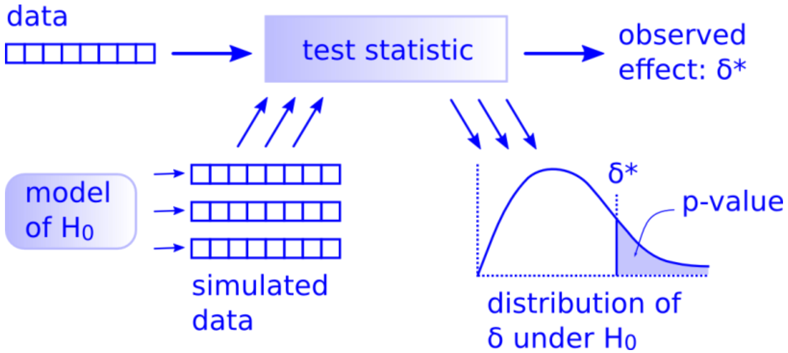
The big picture: There is only one hypothesis test!





- Sample distribution
  - Normal
    - Paired
      - 2 Groups
        - Student's t test for paired samples
      - More Groups
        - ANOVA
    - Unpaired
      - 2 Groups
        - Student's t test for unpaired samples
      - More Groups
        - ANOVA
  - Not Normal
    - Paired
      - 2 Groups
        - Wilcoxon signed rank test
      - More Groups
        - Friedman
    - Unpaired
      - 2 Groups
        - Wilcoxon rank sum test (Mann-Whitney)
      - More Groups
        - Kruskal-Wallis

Just need to follow 5 steps!



data

test statistic

observed effect: δ*

model of H_0

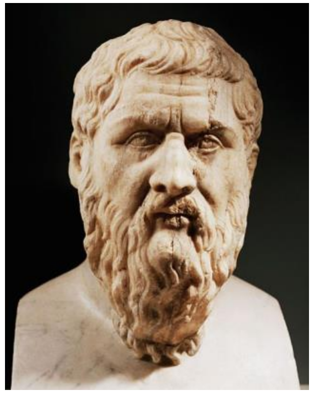
simulated data

δ*

p-value

distribution of δ under H_0

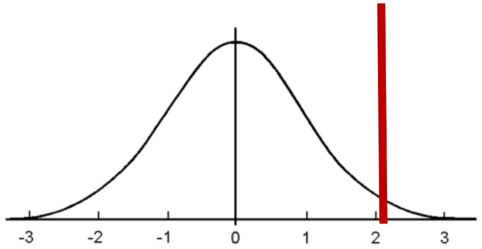
Five steps of hypothesis testing



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

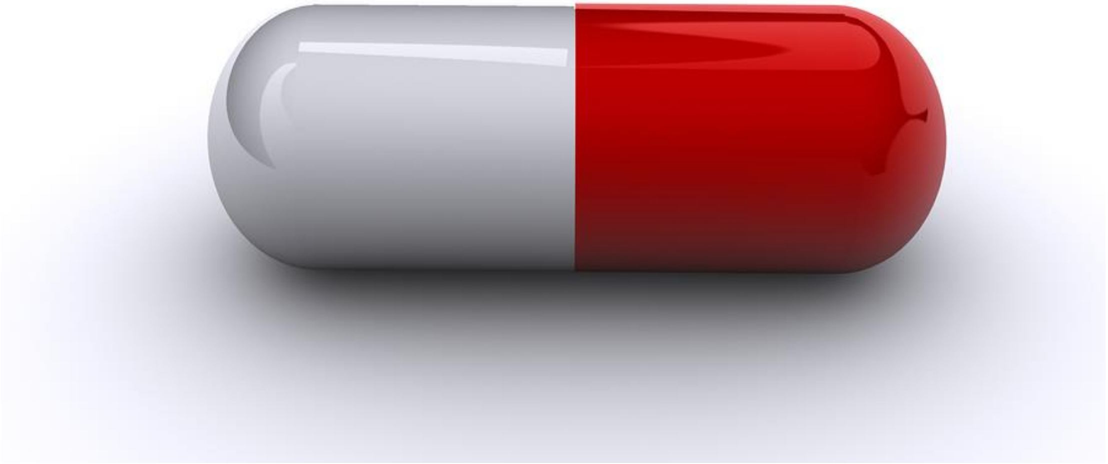
3. Create a distribution of what statistics would look like if Gorgias is right
   - Create the **null distribution** (that is consistent with H_0)



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

Very quick review of randomization test for two means

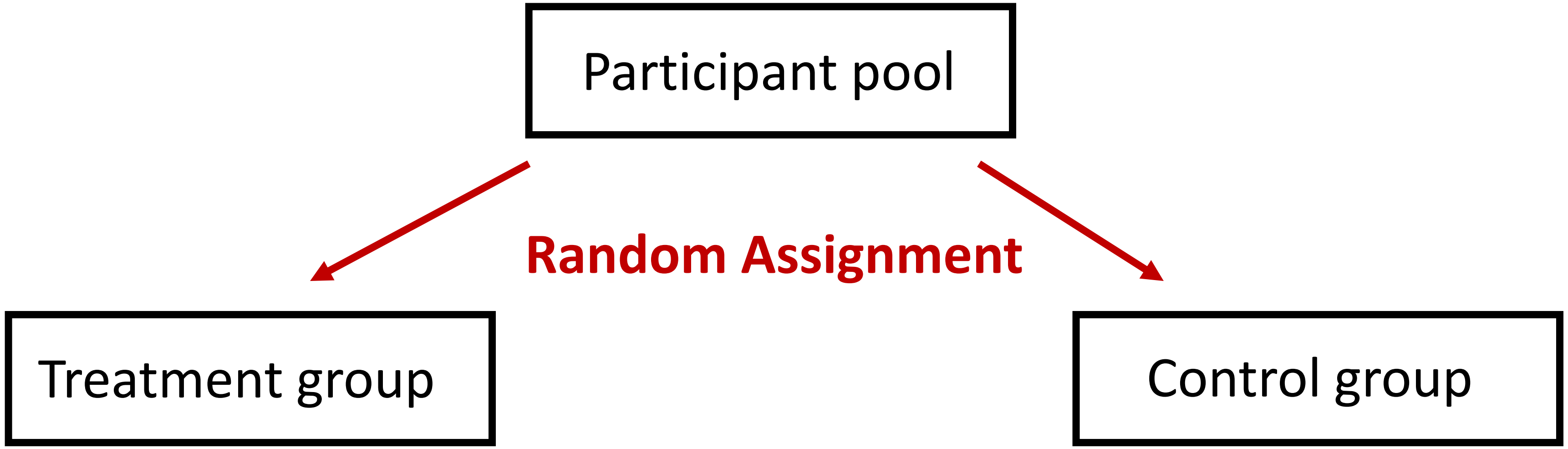


Question: Is this pill effective?

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



Participant pool

Random Assignment

Treatment group

Control group

Hypothesis tests for differences in two group means

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

2. Calculate statistic of interest
   - For randomization/permutation tests we have a choice of the statistic to use

The statistic used before: $\bar{x}_{\text{Effect}} = \bar{x}_{\text{Treatment}} - \bar{x}_{\text{Control}}$

Let’s try Welch's t-statistic instead: $t = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}}$

Does calcium reduce blood pressure?

Treatment data (n = 10):

Begin	107	110	123	129	112	111	107	112	136	102
End	100	114	105	112	115	116	106	102	125	104
Decrease	**7**	**-4**	**18**	**17**	**-3**	**-5**	**1**	**10**	**11**	**-2**

Control data (n = 11):

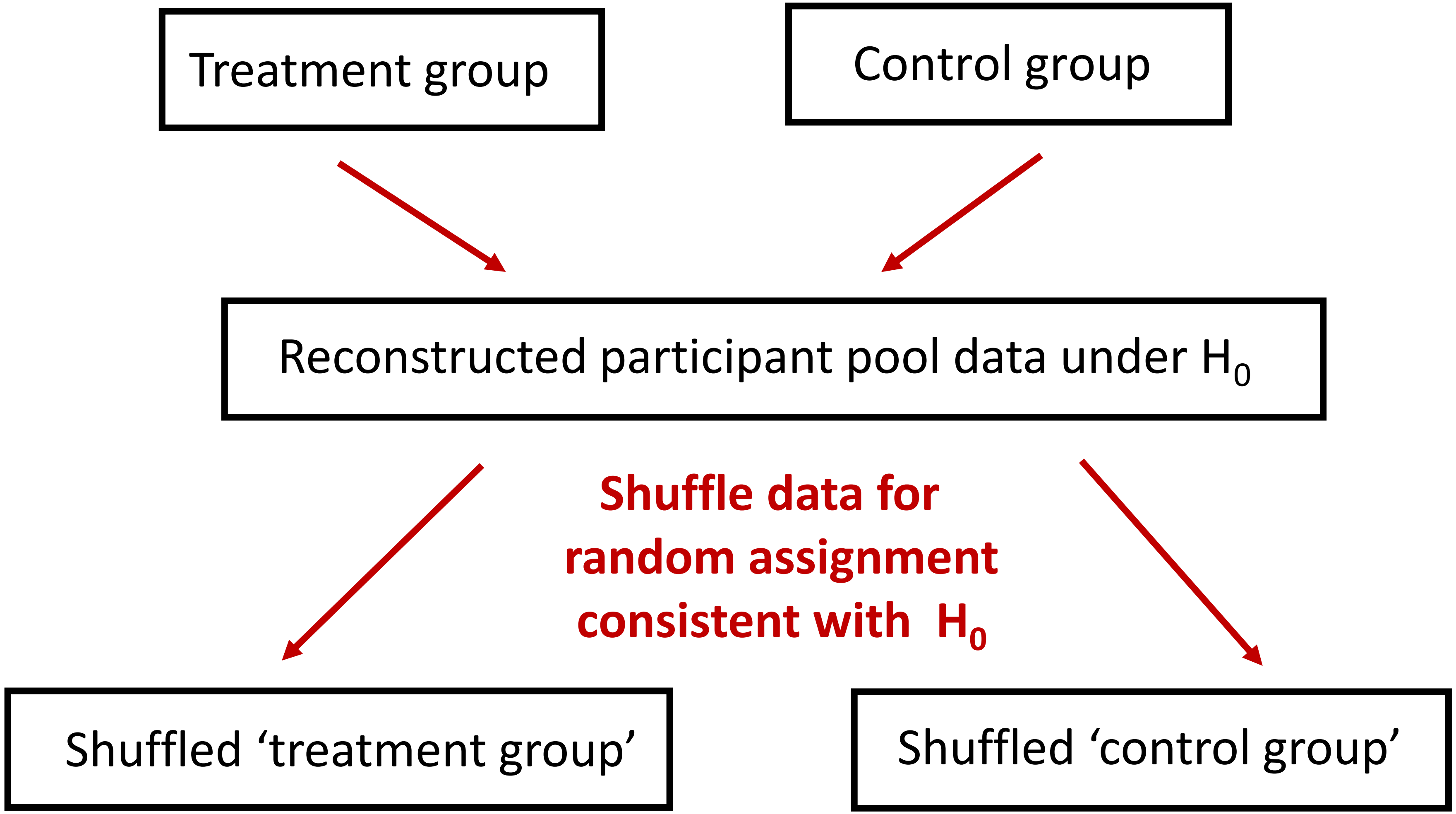
Begin	123	109	112	102	98	114	119	112	110	117	130
End	124	97	113	105	95	119	114	114	121	118	133
Decrease	**-1**	**12**	**-1**	**-3**	**3**	**-5**	**5**	**2**	**-11**	**-1**	**-3**

2. What is the observed statistic of interest?
   - t = 1.604

$$t = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}}$$

3. What is step 3?

3. Create the null distribution!



Treatment group

Control group

Reconstructed participant pool data under H_0

Shuffle data for random assignment consistent with H_0

Shuffled ‘treatment group’

Shuffled ‘control group’

One null distribution statistic: t_{shuff}

Repeat 10,000 times for null distribution

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

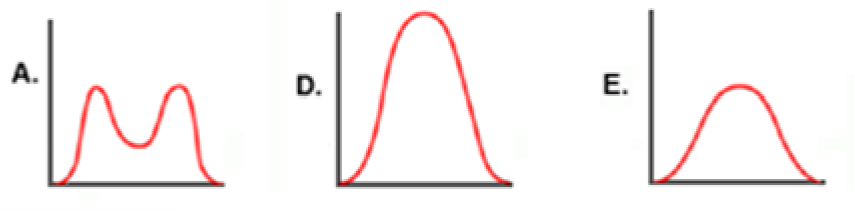
Parametric hypothesis tests

In **parametric hypothesis tests**, the null distribution is given by a density function.

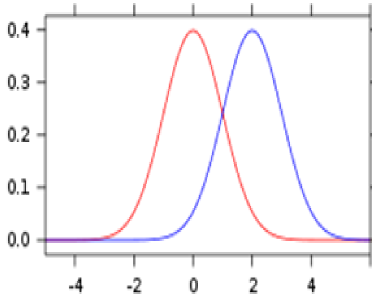
These density functions have a finite set of ***parameters*** that control the shape of these functions

- Hence the name “parametric hypothesis tests”
- Example: the normal density function has two parameters: μ and ***σ***

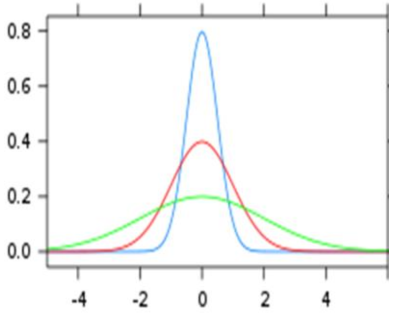
Remember density curves?



Changing μ



Changing σ

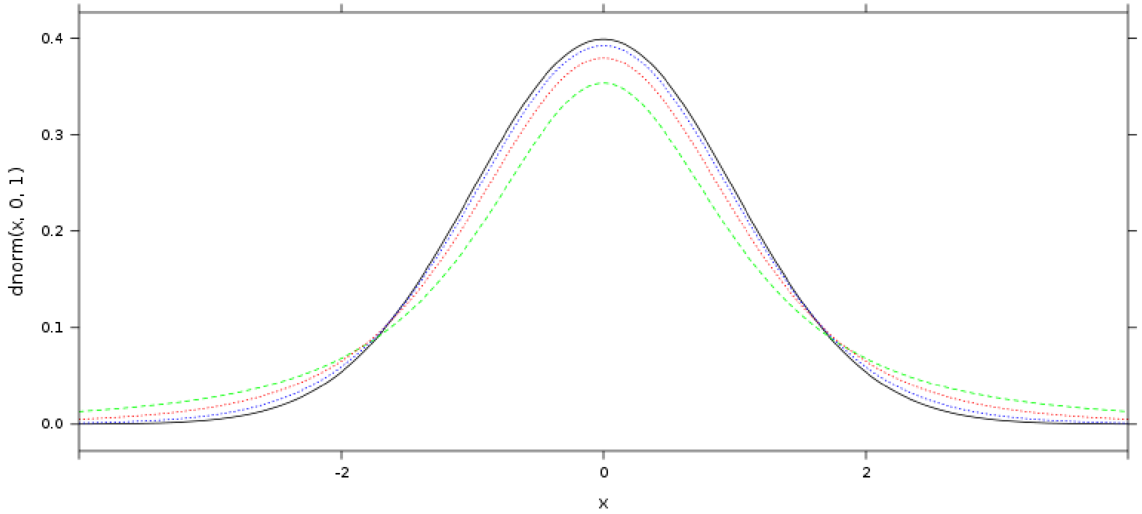


t-distributions

A commonly used density function (distribution) used for statistical inference is the t-distribution

- In R: rt(), dt(), pt() and qt()

t-distributions have one parameter called “degrees of freedom”



df = 2

df = 5

df = 15

N(0, 1)

t-distributions

When using t-distributions for statistical inference, each point in our t-distribution is a t-statistic

- i.e., we use t-distributions as null distributions for hypothesis tests and as sampling distributions when creating confidence intervals

t-statistics are a ratio of:

- The departure of an estimated value from a hypothesized parameter value
- Dived by an estimate of the standard error

$$t = \frac{stat \text{ - } \overset{\theta}{param_0}}{\hat{SE}} \longleftarrow$$

If the SE was known exactly the statistic would be a “z-statistic” that comes from a standard normal distribution

t-tests

t-tests are parametric hypothesis tests where the null distribution is a density function called a t-distribution.

t-tests can be used to test:

- If a mean is equal to a particular value: H_0: $\mu = 7$
- If two means are equal: H_0: $\mu_t = \mu_c$
- If a regression coefficient is equal to a particular value: H_0: $\beta = 2$
- etc.

t-tests for comparing two means

Let’s examine t-tests for comparing **two means**

Step 1: what is the null hypotheses?

Step 2a: What is the numerator of the t-statistic?

$$t = \frac{stat \text{ - } param_0}{\hat{SE}}$$

t-tests for comparing two means

Step 2b: What is the denominator of the t-statistic?

$$t = \frac{stat \text{ - } param_0}{\hat{SE}}$$

Students' t-test assumes the variance in each population is the same, and uses an SE estimate of:

$$\hat{SE}_{\bar{x}_t\text{-}\bar{x}_c} = s_p \cdot \sqrt{\frac{1}{n_t} + \frac{1}{n_c}} \qquad s_p = \sqrt{\frac{\sum_i^{n_t}(x_i - \bar{x}_t)^2 + \sum_j^{n_c}(x_j - \bar{x}_c)^2}{n_t + n_c \text{ - } 1}}$$

Welch's t-test does **not** assume that the variance in each population is the same and uses an estimate of:

$$\hat{SE}_{\bar{x}_t\text{-}\bar{x}_c} = \sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}$$

t-tests for comparing two means

Question: which statistic/test is better to use?

Students' t-test assumes the variance in each population is the same, and uses an SE estimate of:

$$t = \frac{\bar{x}_t \text{ - } \bar{x}_t}{s_p \cdot \sqrt{\frac{1}{n_t} + \frac{1}{n_c}}} \qquad s_p = \sqrt{\frac{\sum_i^{n_t}(x_i - \bar{x}_c)^2 + \sum_j^{n_c}(x_j - \bar{x}_c)^2}{n_t + n_c \text{ - } 1}}$$

Welch's t-test does **not** assume that the variance in each population is the same and uses an estimate of:

$$\hat{SE}_{\bar{x}_t\text{-}\bar{x}_c} = \sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}} \qquad t = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}}$$

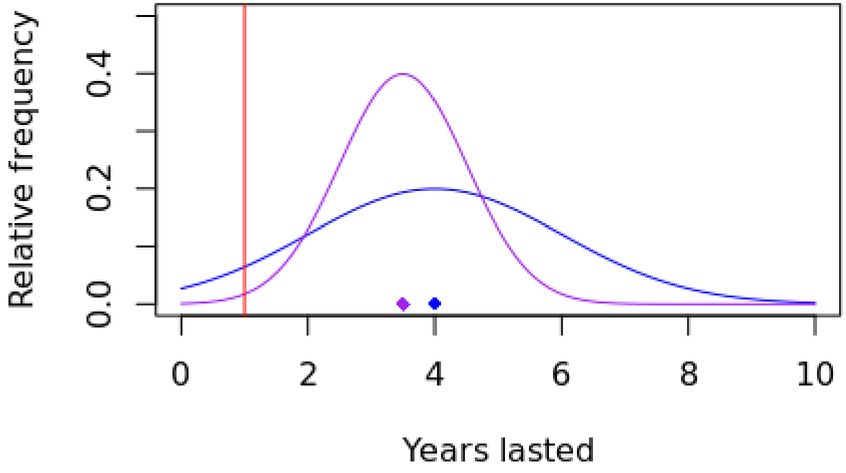
t-tests for comparing two means

Question: which statistic/test is better to use?

However, we need to be careful with the decisions we make based on differences of means when there are unequal variances.

E.g., Which car battery company produces better batteries in terms of how long they last?

- Company A: μ = 4 years, σ = 2 years
- Company B: μ = 3.5 years, σ = 1 years



- Company A: 7% fail within a year
- Company B: 0.6% fail with a year

Does calcium reduce blood pressure?

Treatment data (n = 10):

Begin	107	110	123	129	112	111	107	112	136	102
End	100	114	105	112	115	116	106	102	125	104
Decrease	**7**	**-4**	**18**	**17**	**-3**	**-5**	**1**	**10**	**11**	**-2**

Control data (n = 11):

Begin	123	109	112	102	98	114	119	112	110	117	130
End	124	97	113	105	95	119	114	114	121	118	133
Decrease	**-1**	**12**	**-1**	**-3**	**3**	**-5**	**5**	**2**	**-11**	**-1**	**-3**

2. What is the observed statistic of interest?
   - t = 1.604

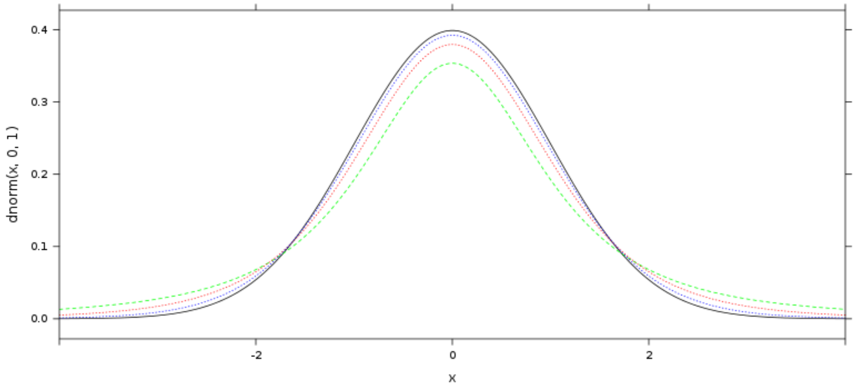
$$t = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}}$$

3. What is the null distribution?
   - What additional piece of information do we need to create it?

t-tests for comparing two means

When using a t-distribution to compare two means, a conservative estimate of the degrees of freedom is the minimum of the two samples sizes, n_t and n_c, minus 1

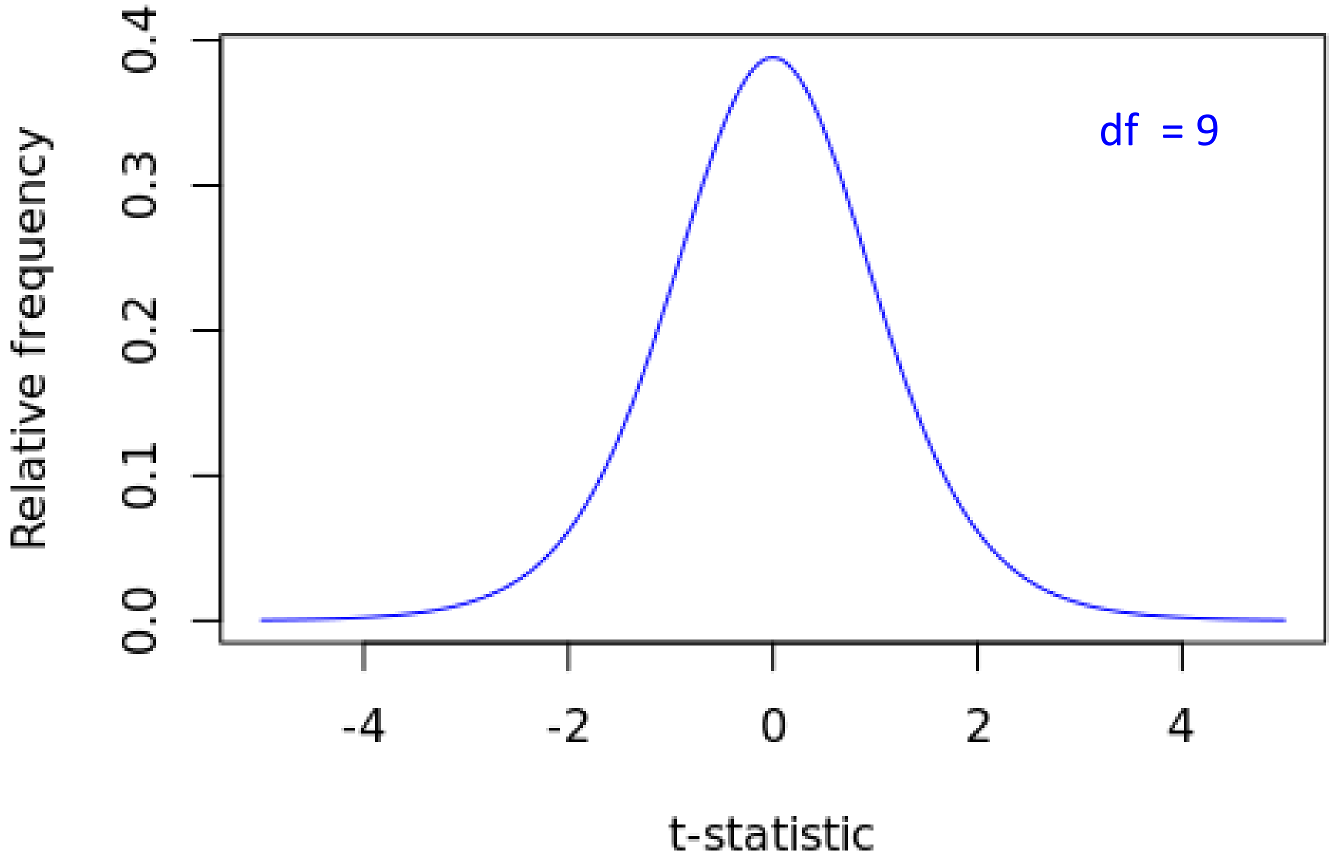
- df = min(n_t, n_c) – 1



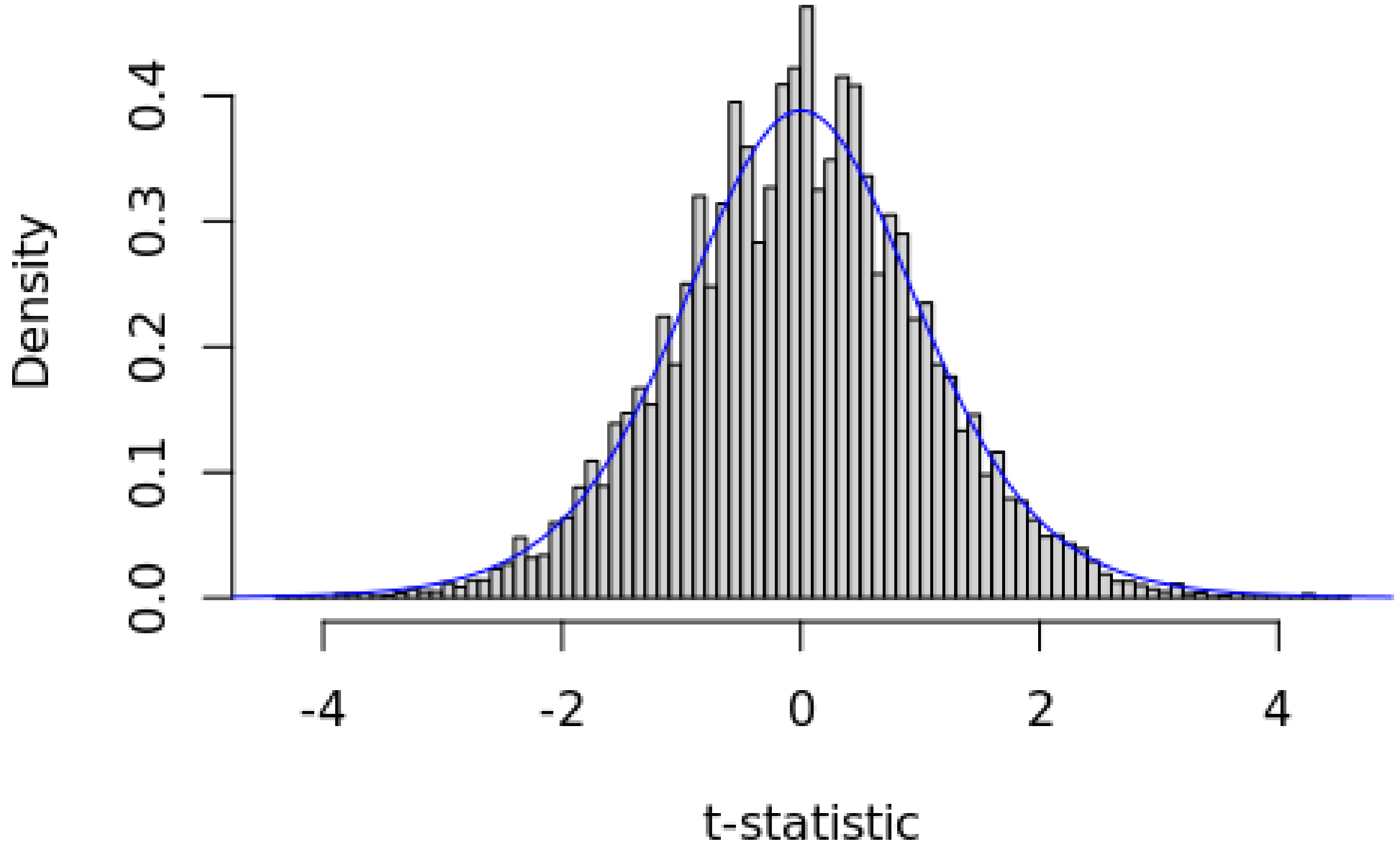
Q: For the calcium study we had 10 people in the control group and 11 people in the treatment group so the degrees of freedom parameter is?

- 9

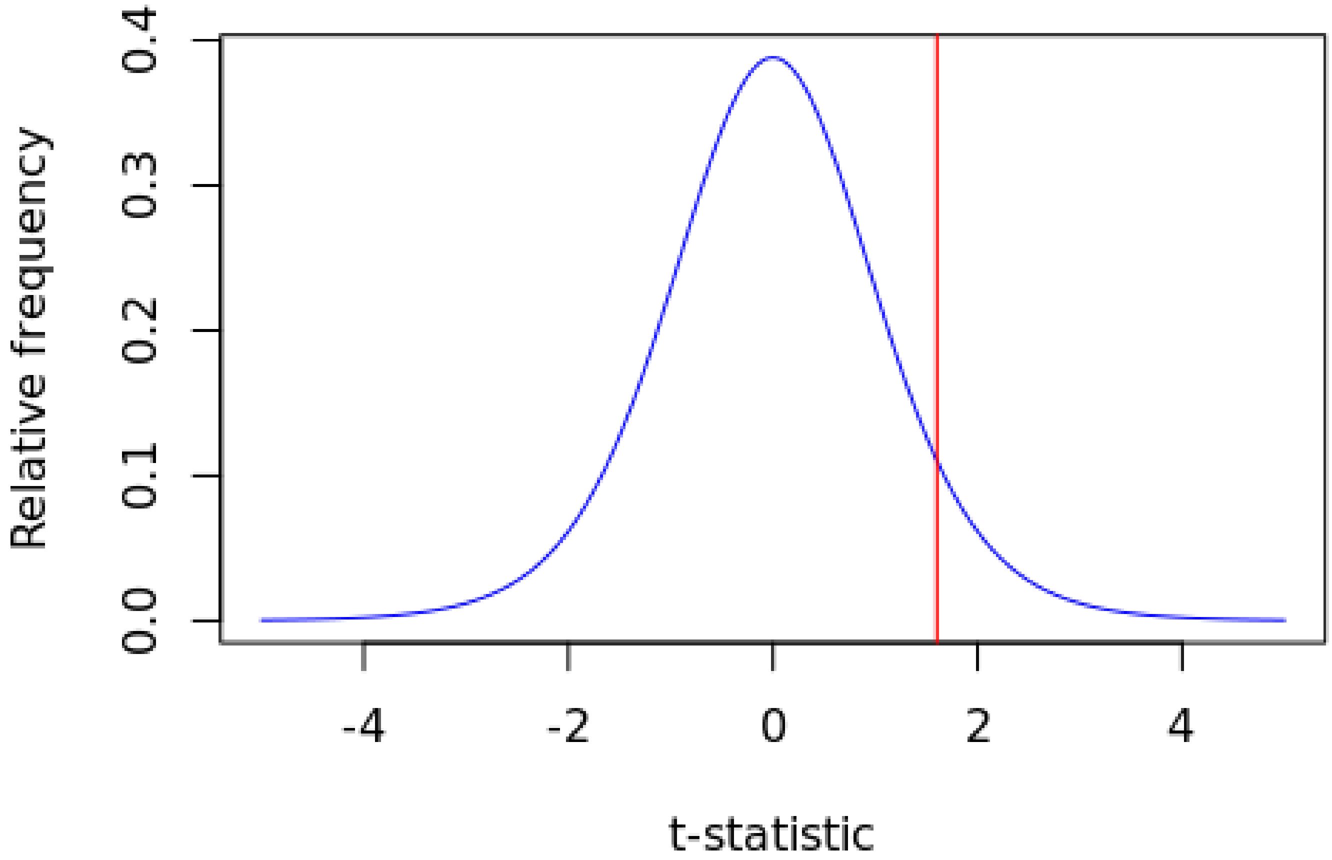
Step 3: Null t-distribution



Step 3: parametric vs. randomization distributions



Step 4-5: p-value and conclusion

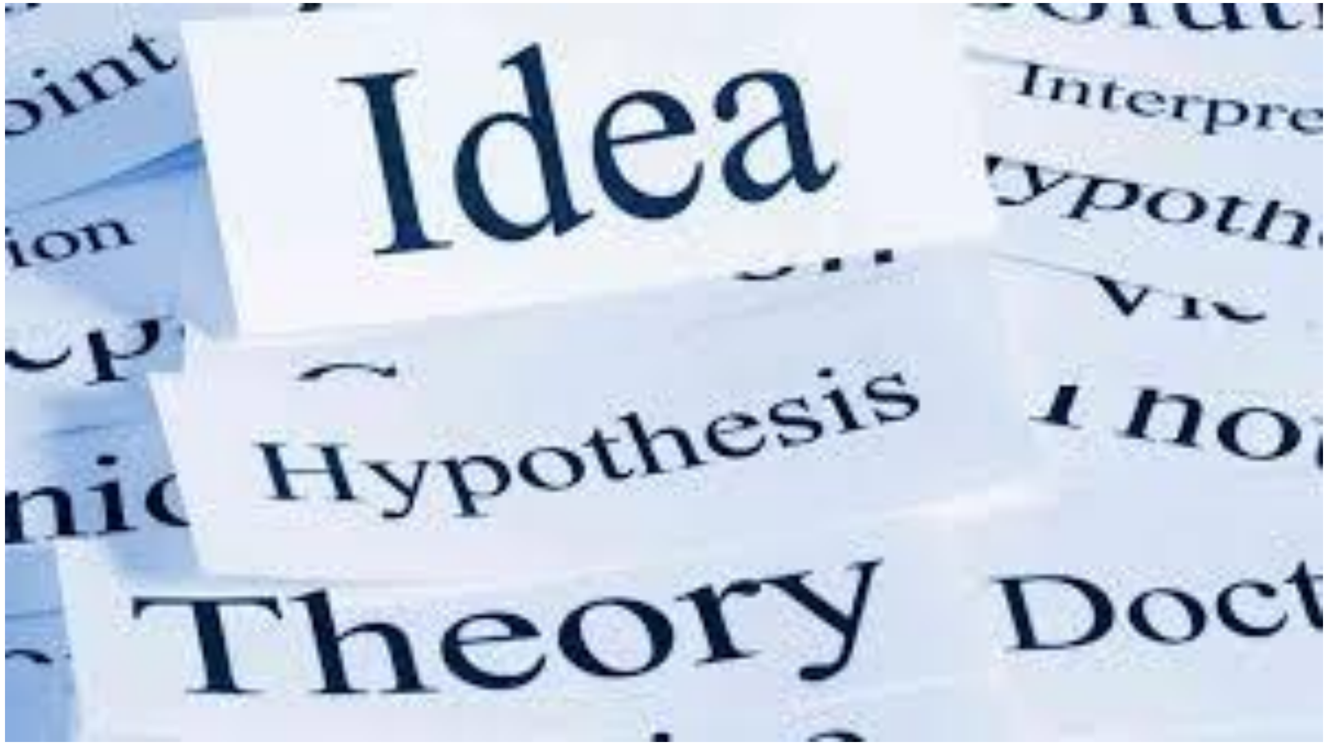


p-value = 0.072

Conclusion?

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)

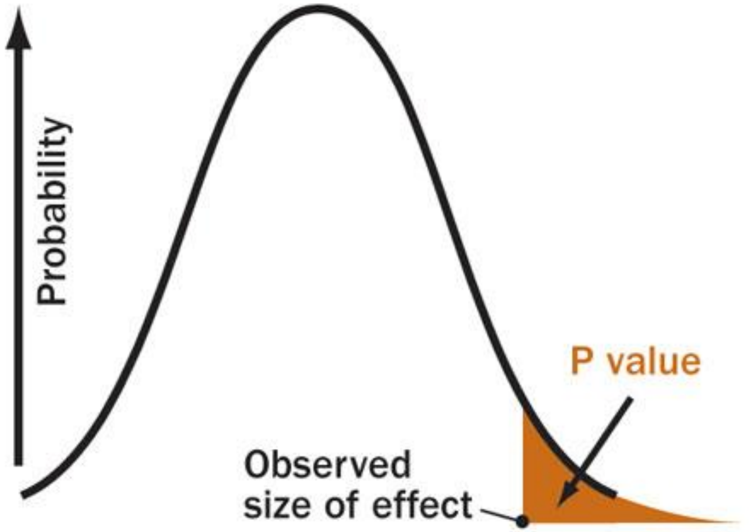
Neyman (1894-1981)

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: tells the experimenter “what results to ignore”



Neyman-Pearson null hypothesis testing

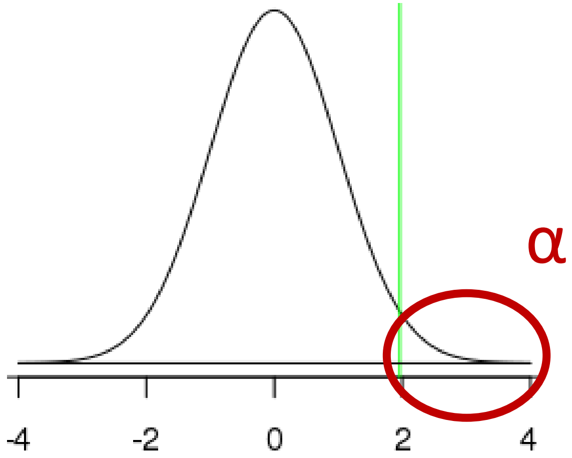
Makes ***a formal decision*** in statistical tests

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

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

Null distribution



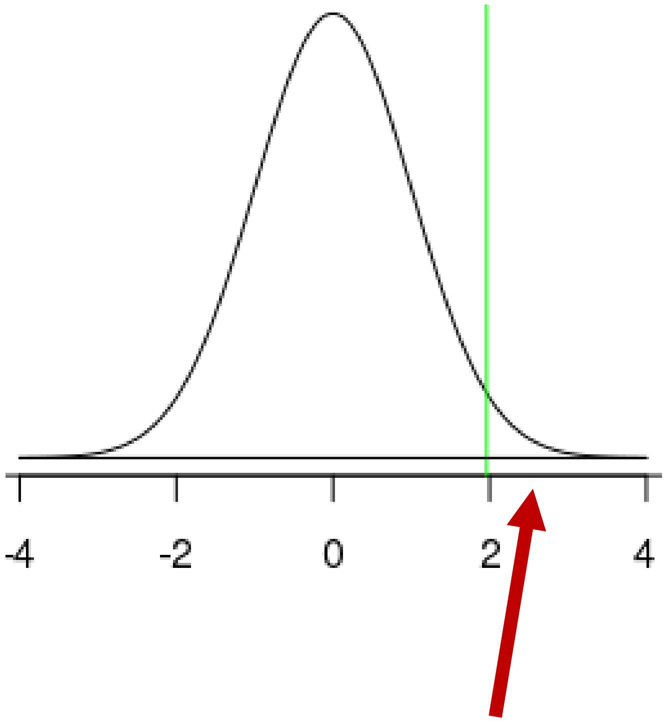
Neyman-Pearson frequentist logic

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

If Neyman-Pearson null hypothesis testing paradigm was followed perfectly, then only ~5% of all published research findings should be wrong (for α = 0.05)

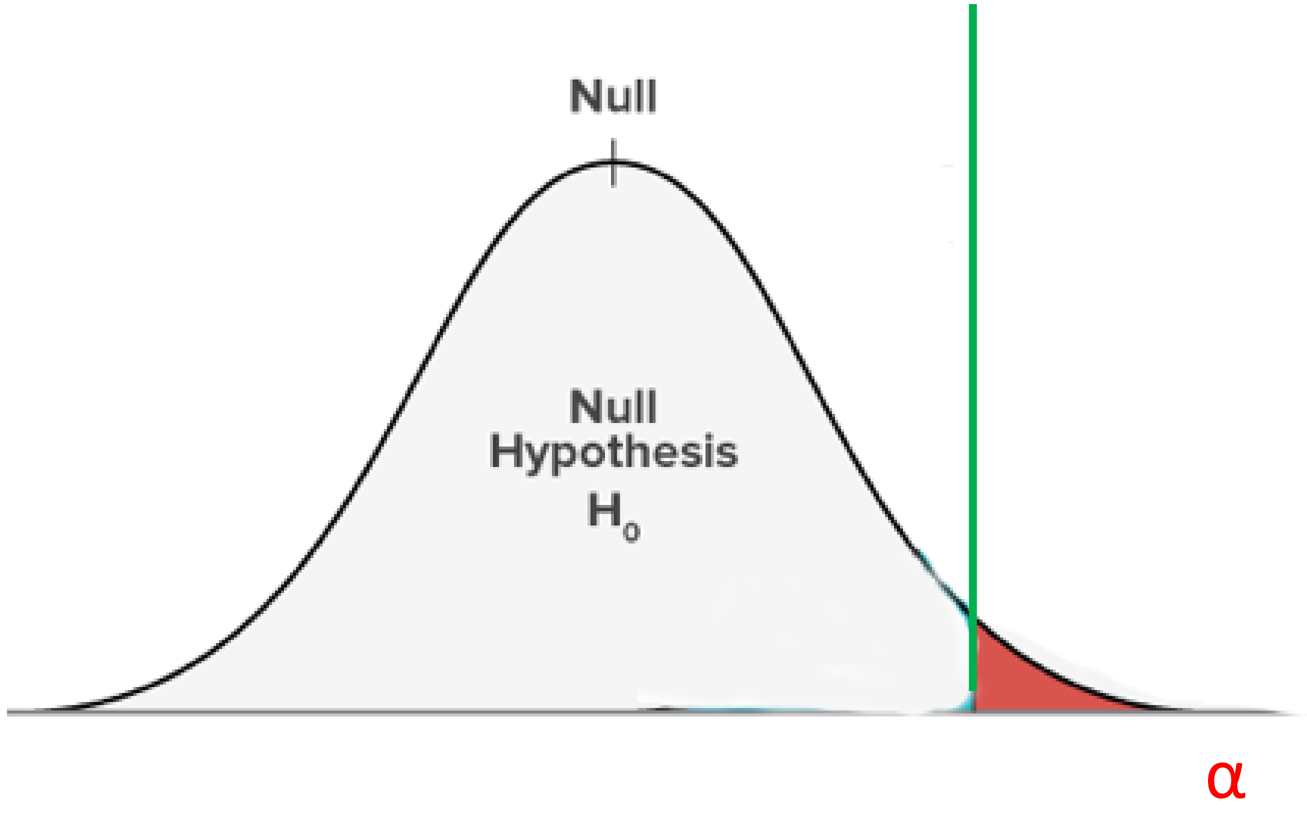
- i.e., we would only make type I errors 5% of the time

Null distribution

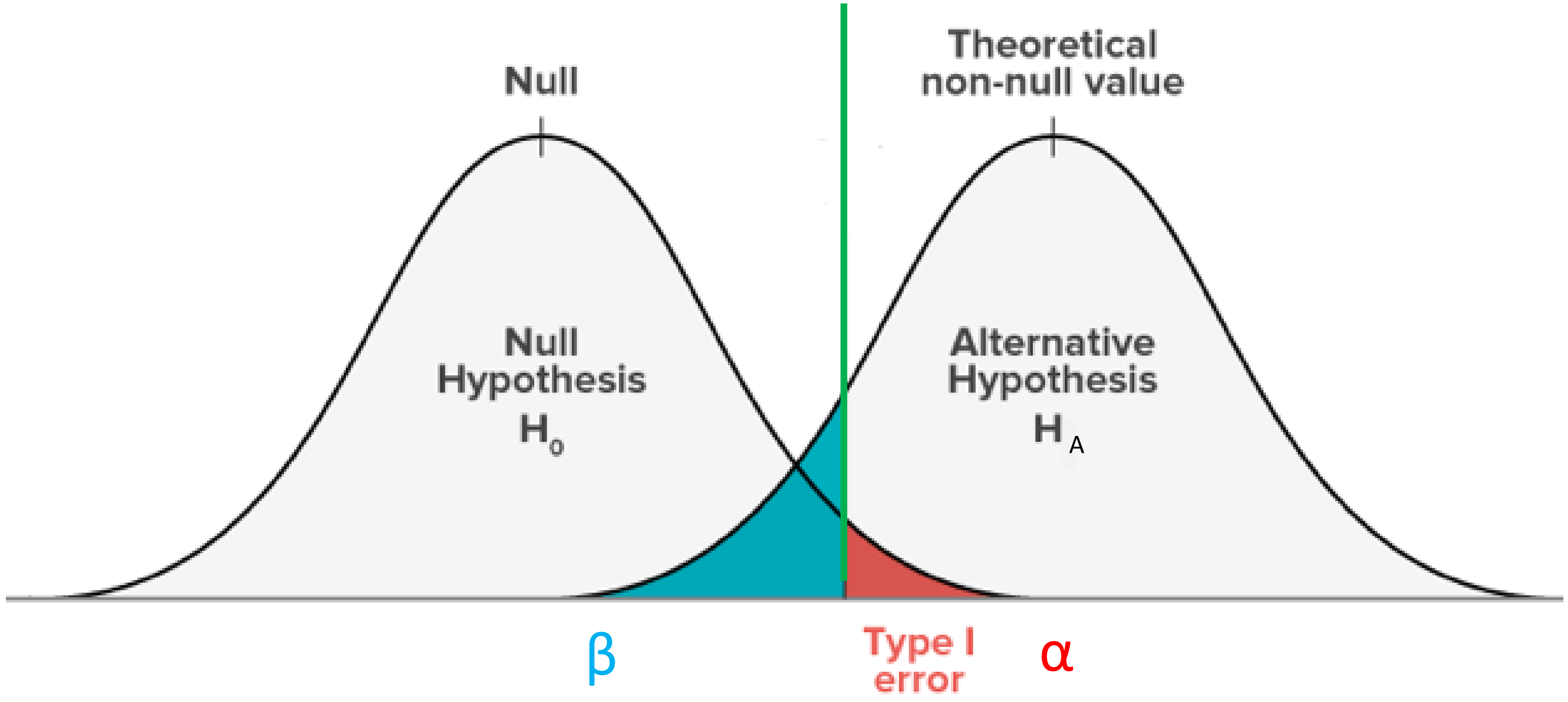


The null distribution is true but statistic landed here

Neyman-Pearson Frequentist logic



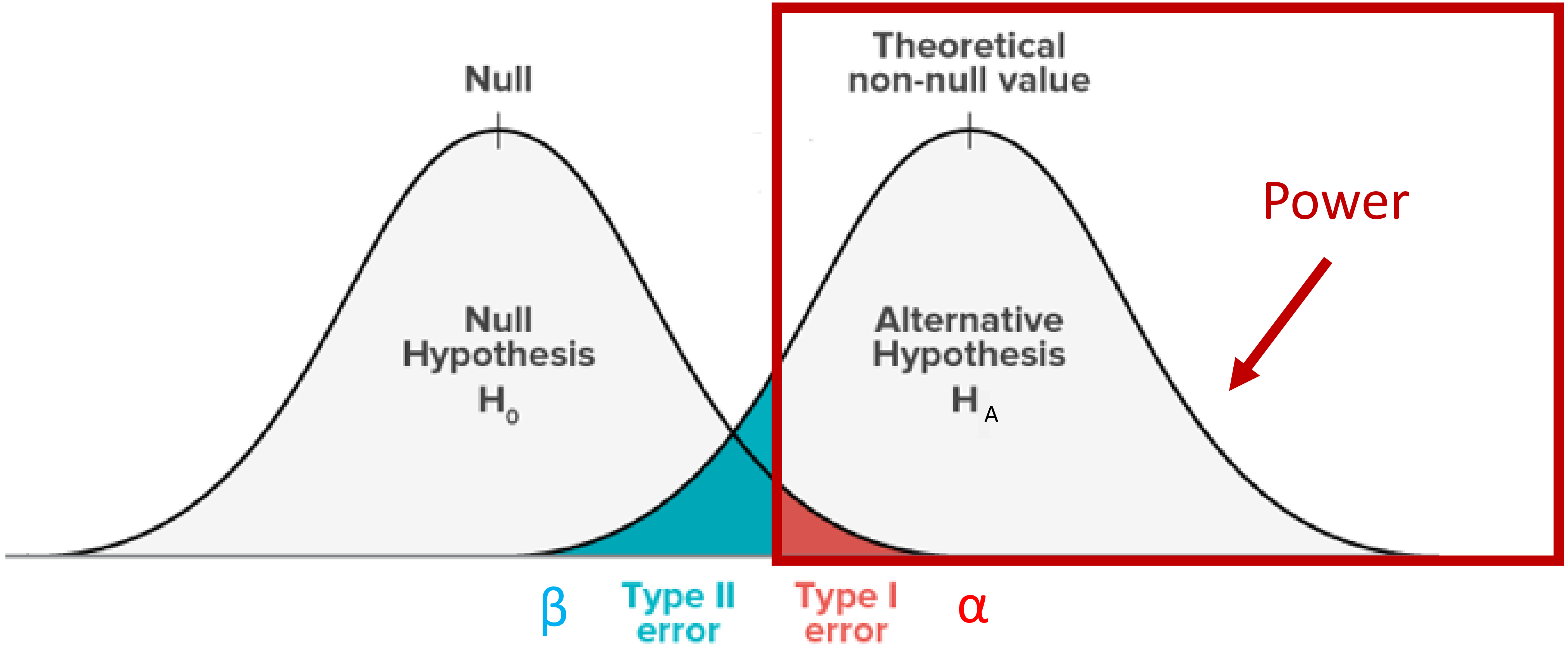
Neyman-Pearson Frequentist logic



Type 2 error: incorrectly rejecting failing to reject H_0 when it is false

- The rate at which we make type 2 errors is often denoted with the symbol β

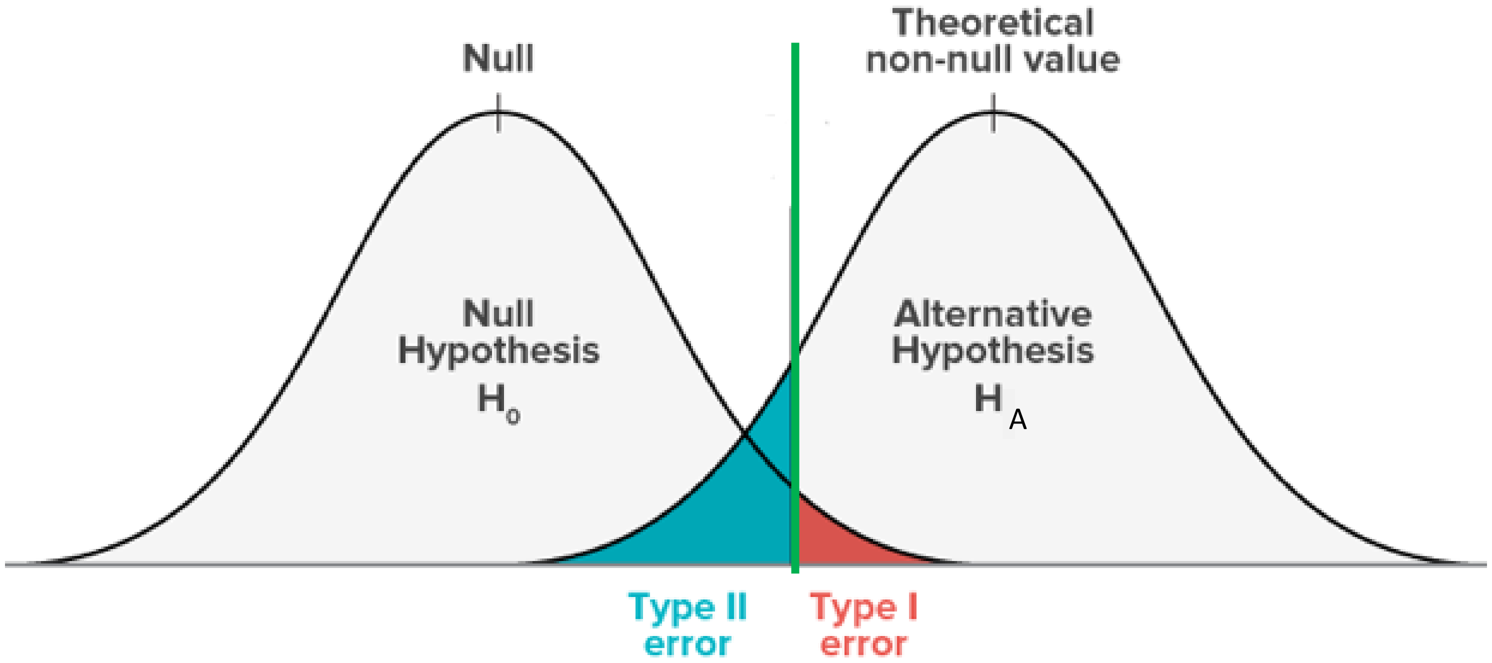
Neyman-Pearson Frequentist logic



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



	Reject H_0	Do not reject H_0
H_0 is true	Type I error (α) (false positive)	No error
H_A is true **(H_o is false)**	No error	Type II error (β) (false negative)

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 true:
    - Joy can smell Parkinson's disease, Lawyers are left-handed at a higher rates than the general population, Calcium is good 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 true:
    - Joy can smell Parkinson's disease, Lawyers are left-handed at a higher rates than the general population, Calcium is good 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 1 errors

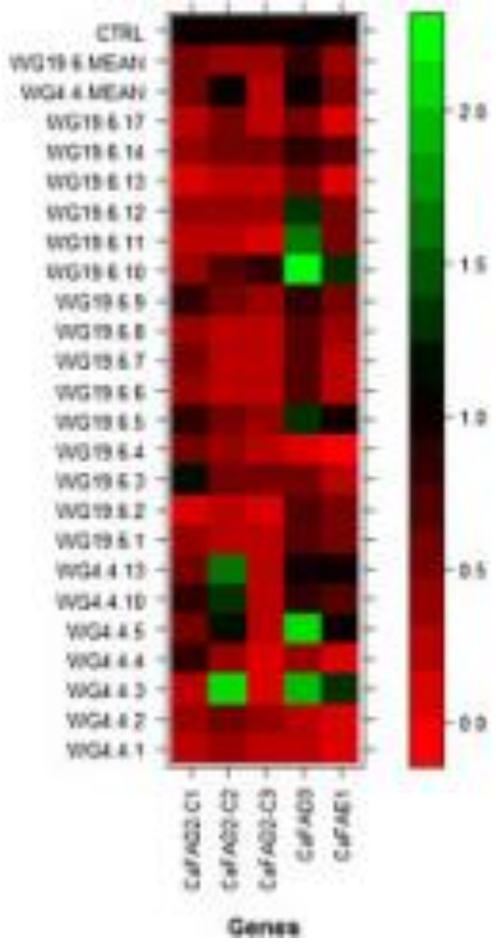
Genes and leukemia example

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 1 errors would be expected?



Genes and leukemia example

There are methods that try to correct for running multiple hypothesis tests

The ***Bonferroni correction*** is one way that controls the probability of ***any*** hypothesis test giving a type 1 error

- i.e., controls the familywise error rate (no type 1 errors for any of the tests run)

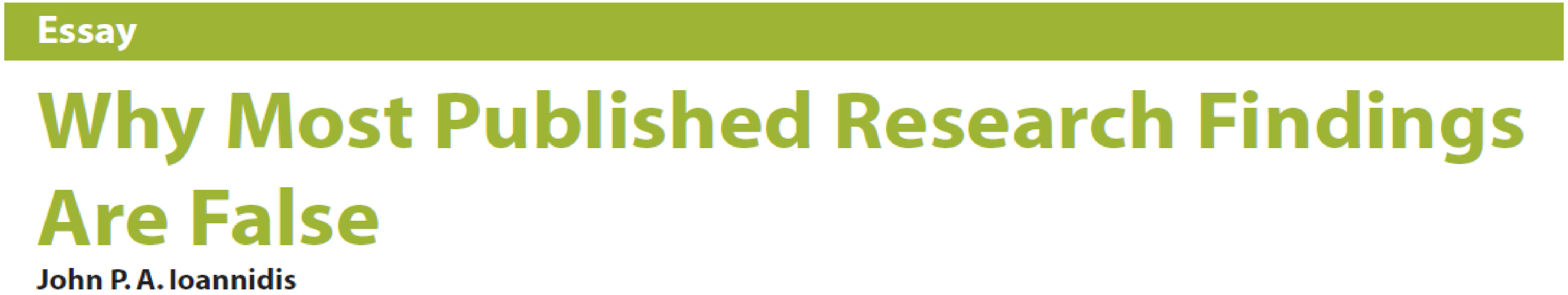
It works by dividing the initial α level by the number of tests run

- E.g., $\alpha = 0.05/7129 = 0.000007$
- All p-values need to be below this level to be considered statistically significant
- This can lead to many type 2 errors (Type 2 error: failure to reject H_0 when it is false)

The problem of multiple testing

For α = 0.05, ~5% of all published research findings should be wrong

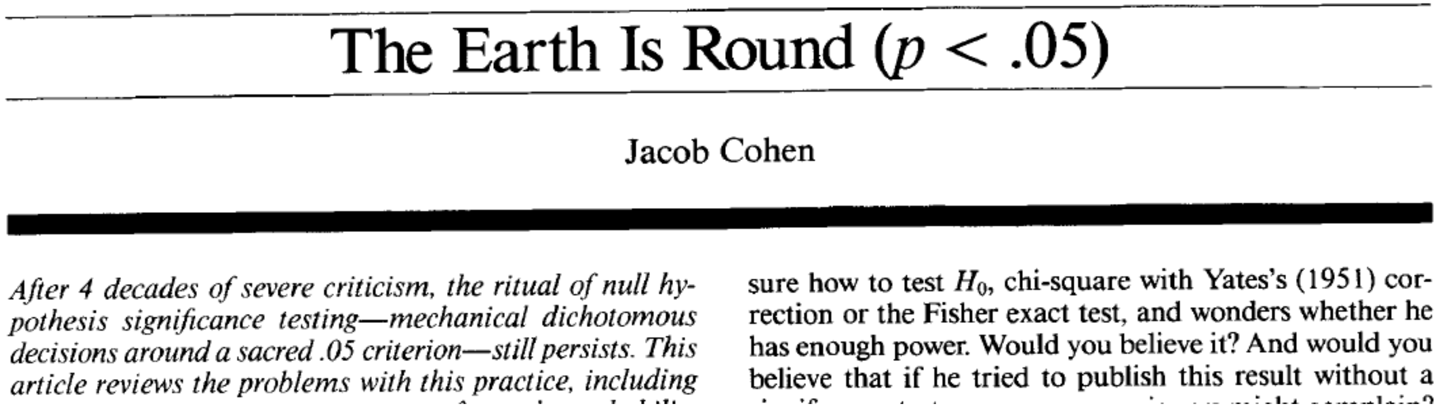
Publication bias (file drawer effect): Generally positive results are more likely to be published, so if you read the literature, the number of incorrect results (type 1 errors) will be greater than 5%.



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!