Hypothesis tests for more than two means

	5	3	2		7			8
6		1	5					2
2			9	1	3		5	
7	1	4	6	9	2			
	2						6	
			4	5	1	2	9	7
	6		3	2	5			9
1					6	3		4
8			1		9	6	7	

Overview

Worksheet 9 questions

Review of hypothesis testing for two means

Hypothesis tests for more than 2 means

Theories and issues with hypothesis tests

Final project: analyze your own data set

Final project report: a 5-10 page R Markdown document that contains:

1. Background information:
   - What question you will answer and why it is interesting
   - Where you got the data, and any prior analyses
2. Descriptive plots
3. At least one hypothesis test
4. At least one confidence interval
5. A conclusion and reflection

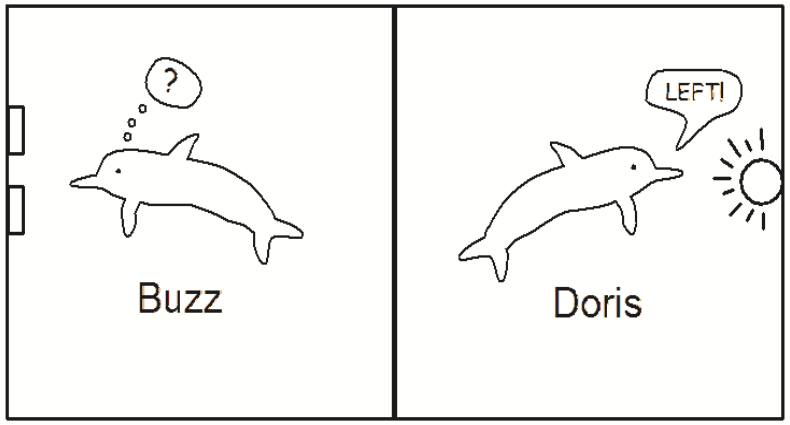
A one paragraph final project proposal is due on Tuesday November 20th

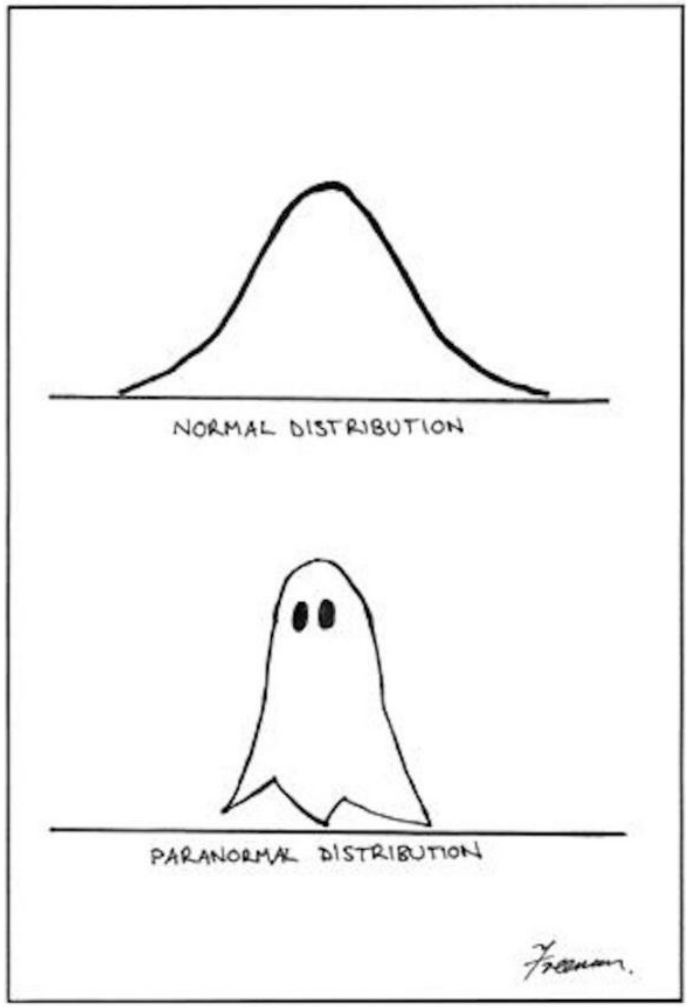
- What question you will answer
- Where you will get the data

Worksheet 9 questions?

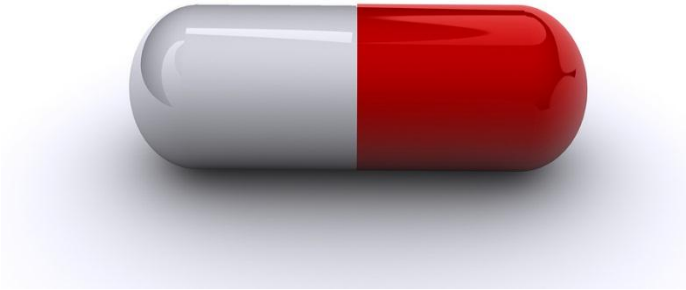
How did it go?

Review: hypothesis tests





[Movie 1] [Movie 2]

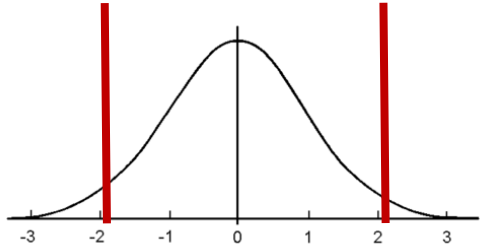


Quiz: please write down the 5 steps of hypothesis testing

1. State H0 and HA
   - Assume innocence: H0 is true

2. Calculate the observed statistic
   - Gather evidence

3. Create the null distribution
   - A distribution of what evidence would look like if H_0 is true



4. Calculate the p-value
   - Assess the probability that the observed evidence would come from the null 'innocent' distribution

5. Assess whether the results are statistically significant
   - Make a verdict: innocent or guilty

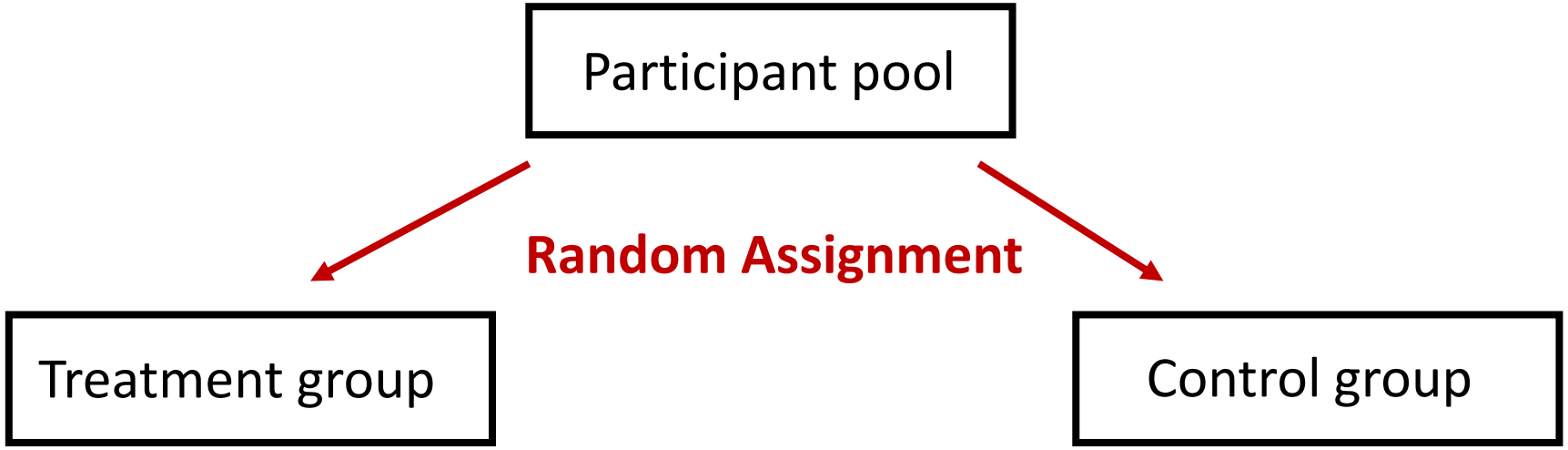
Hypothesis tests for comparing two means

Question: Can we find out the ***Truth*** of whether the 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



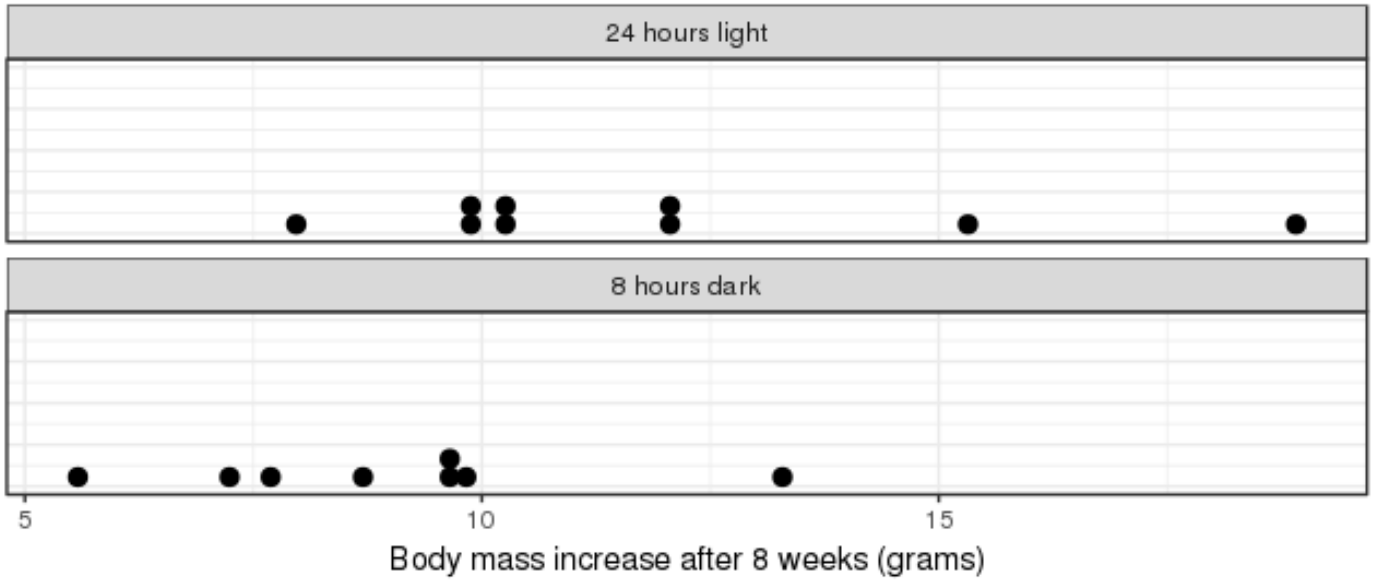
Do mice who eat late at night get fat?

A study by Fonken et al, 2010, wanted to examine whether more weight was gained by mice who could eat late at night

Mice were randomly divided into 2 groups:

- Dark condition: 8 mice were given 8 hours of darkness at night (when they couldn't eat)
- Light condition: 9 were constantly exposed to light for 24 hours (so they could always eat)

What's a good first thing to do when analyzing data?



Hypothesis tests for differences in two group means

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

2. Calculate statistic of interest
   - $\bar{x}_{effect} = \bar{x}_{Dark} - \bar{x}_{Light}$

Do mice who eat late at night get fat?

You can get the data from:

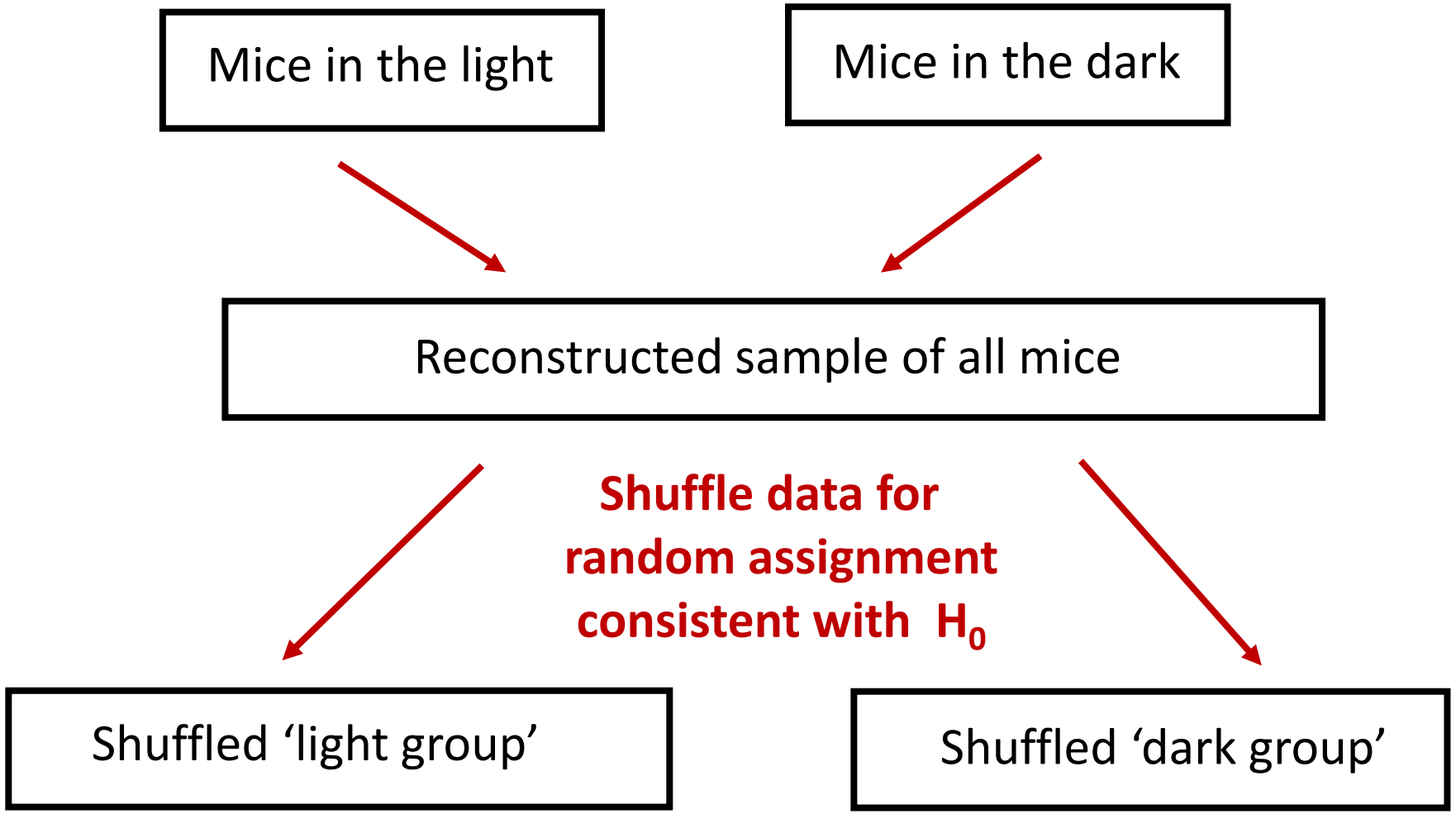
```
>  load("/home/shared/intro_stats/cs206_data/mice.Rda")

> dark_BM_increase      # length(dark_BM_increase)
> light_BM_increase     # length(light_BM_increase)
```

Can you calculate the observed statistic (step 2)?

```
    > obs_stat <- mean(light_BM_increase) - mean(dark_BM_increase)
```

What's next?

3. Create the null distribution!



One null distribution statistic: $\bar{x}_{Shuff\_Dark} - \bar{x}_{Shuff\_Light}$

Do mice who eat late at night get fat?

What is the first thing we need to do for creating the null distribution?

```
combo_data <- c(light_BM_increase, dark_BM_increase)
```

How do we create one point in our null distribution?

```
# shuffle the data
shuff_data  <- sample(combo_data)

# create fake light and dark data
shuff_light <- shuff_data[1:9]
shuff_dark <- shuff_data[10:17]

# compute fake statistic
mean(shuff_light) - mean(shuff_dark)
```

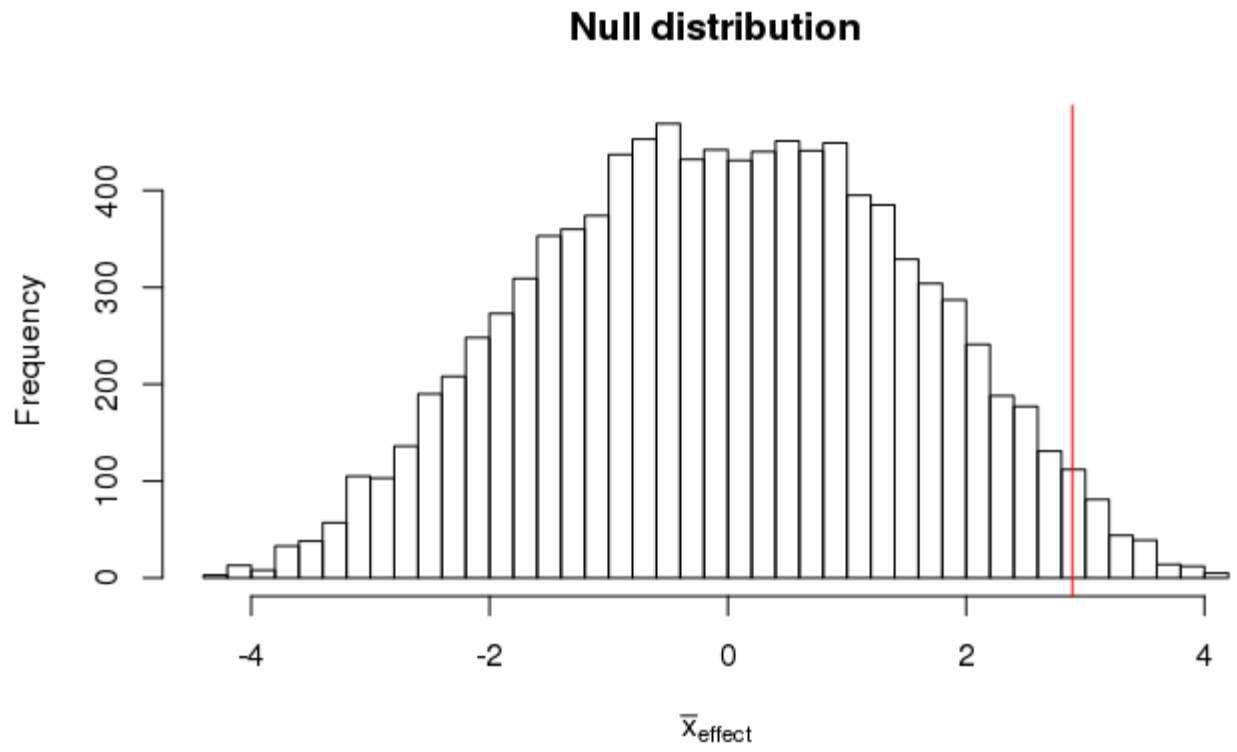
Do mice who eat late at night get fat?

How do we create a full null distribution?

```
null_dist <- NULL
for (i in 1:10000) {

        shuff_data <- sample(combo_data)
        shuff_light <- shuff_data[1:9]
        shuff_dark <- shuff_data[10:17]
        null_dist[i] <- mean(shuff_light) - mean(shuff_dark)

}
```

Do mice who eat late at night get fat?

Plot the null distribution: hist(null_dist, nclass = 50)



What do we do next?

Do mice who eat late at night get fat?

Get the p-value

```
p_val <- sum(null_dist >= obs_stat)/10000
```

p-value = 0.02

Comparing more than two means

A group of Hope College students wanted to see if there was an association between a student's major and the time it takes to complete a small Sudoku-like puzzle

	5	3	2		7			8
6		1	5					2
2			9	1	3		5	
7	1	4	6	9	2			
	2						6	
			4	5	1	2	9	7
	6		3	2	5			9
1					6	3		4
8			1		9	6	7	

Comparing more than two means

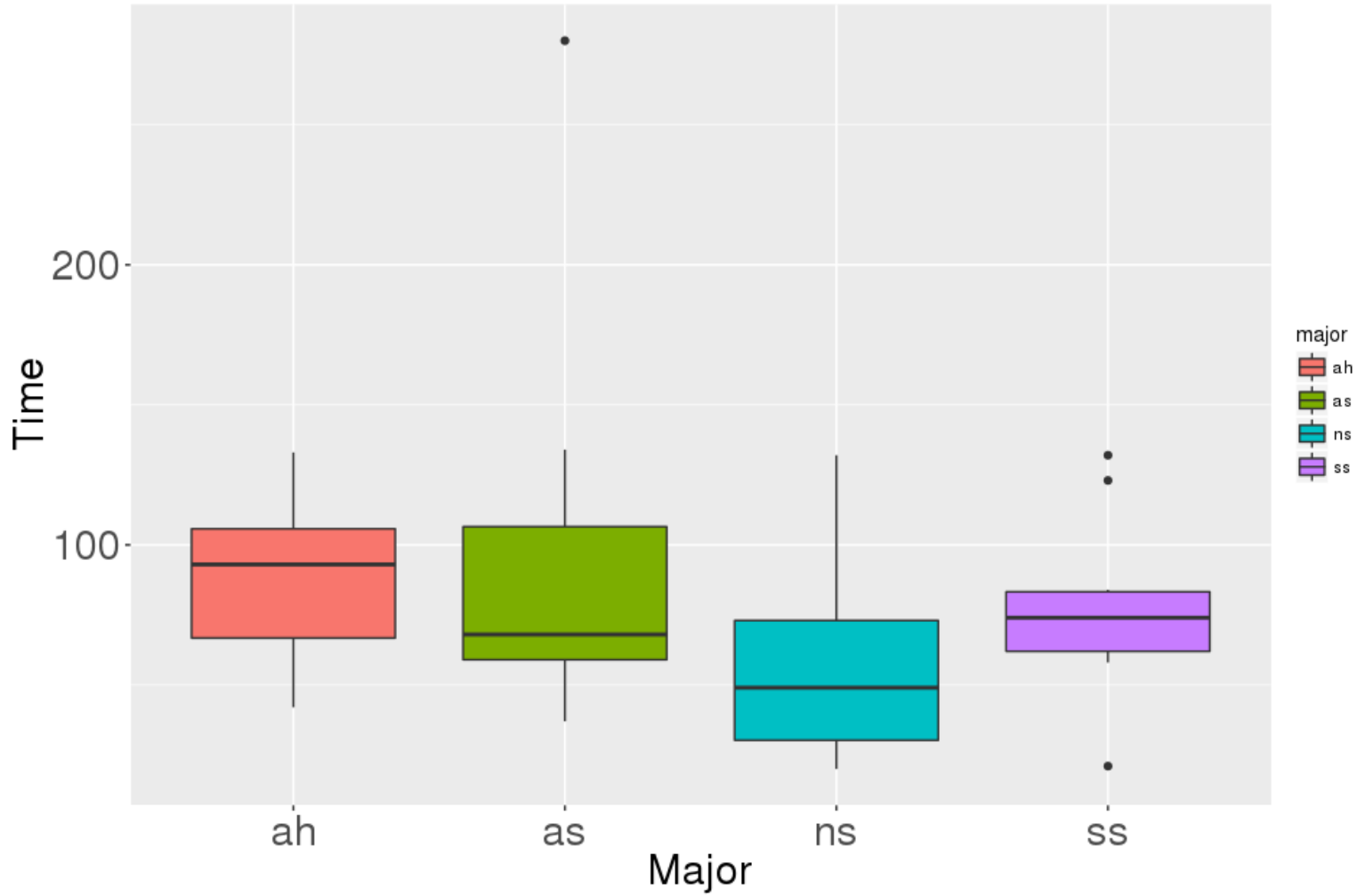
A group of Hope College students wanted to see if there was an association between a student's major and the time it takes to complete a small Sudoku-like puzzle

They grouped majors into four categories

- Applied science (as)
- Natural science (ns)
- Social science (ss)
- Arts/humanities (ah)

What is the first thing to do to analyze the data?

Step 0: Plot of completion time by major



What should we do next?

Sudoku by field

1. State the null and alternative hypotheses!

$\mathbf{H_0}$: $\mu_{as} = \mu_{ns} = \mu_{ss} = \mu_{ah}$

$\mathbf{H_A}$: $\mu_i \neq \mu_j$ for one pair of fields of study

What should we do next?

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):

   $(|\bar{x}_{as} - \bar{x}_{ns}| + |\bar{x}_{as} - \bar{x}_{ss}| + |\bar{x}_{as} - \bar{x}_{ah}| + |\bar{x}_{ns} - \bar{x}_{ss}| + |\bar{x}_{ns} - \bar{x}_{ah}| + |\bar{x}_{ss} - \bar{x}_{ah}|)/6$

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

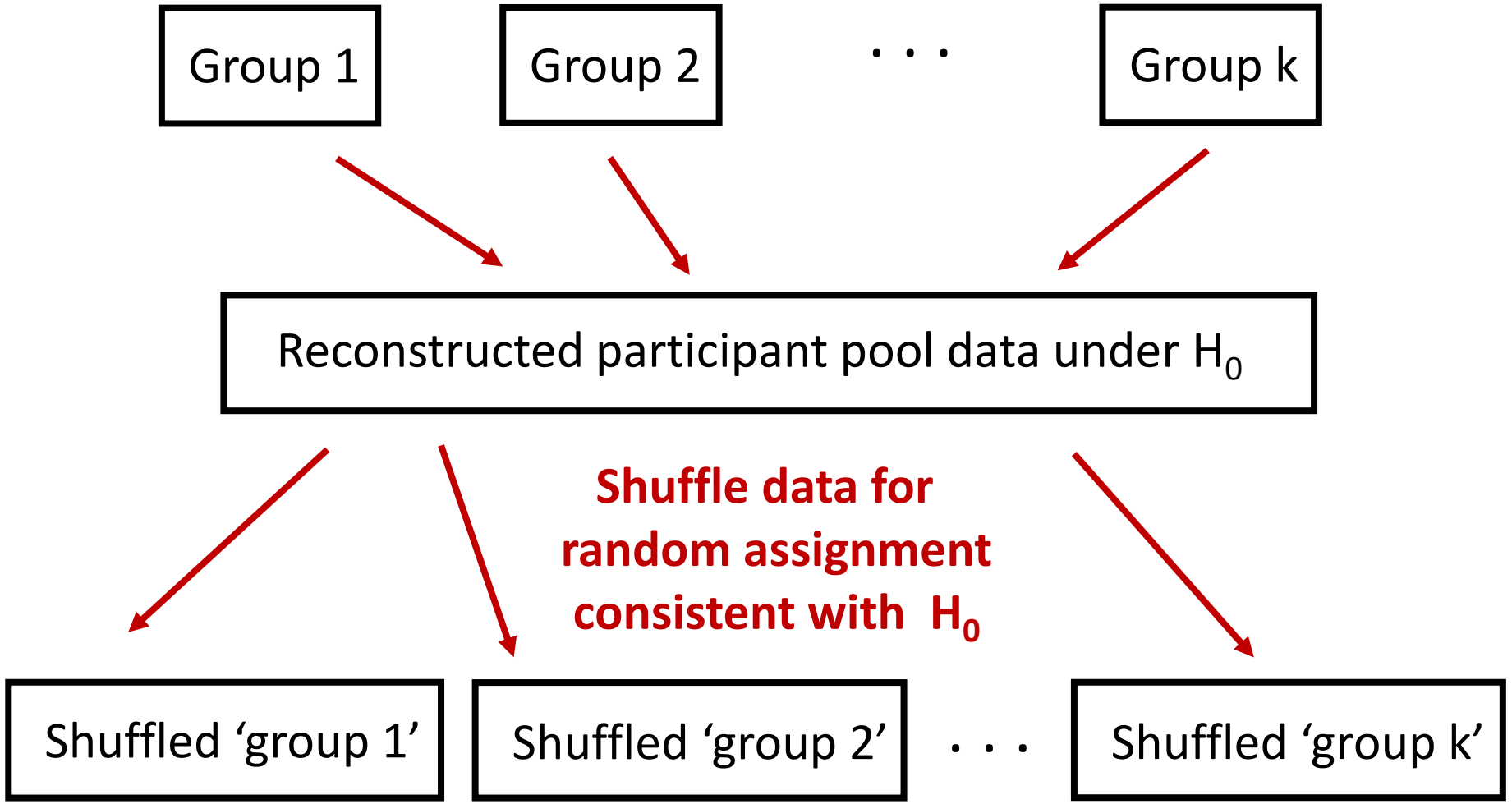
Mean absolute difference (MAD):

$$(|\bar{x}_{as} - \bar{x}_{ns}| + |\bar{x}_{as} - \bar{x}_{ss}| + |\bar{x}_{as} - \bar{x}_{ah}| + |\bar{x}_{ns} - \bar{x}_{ss}| + |\bar{x}_{ns} - \bar{x}_{ah}| + |\bar{x}_{ss} - \bar{x}_{ah}|)/6$$

Observed statistic value = 13.92

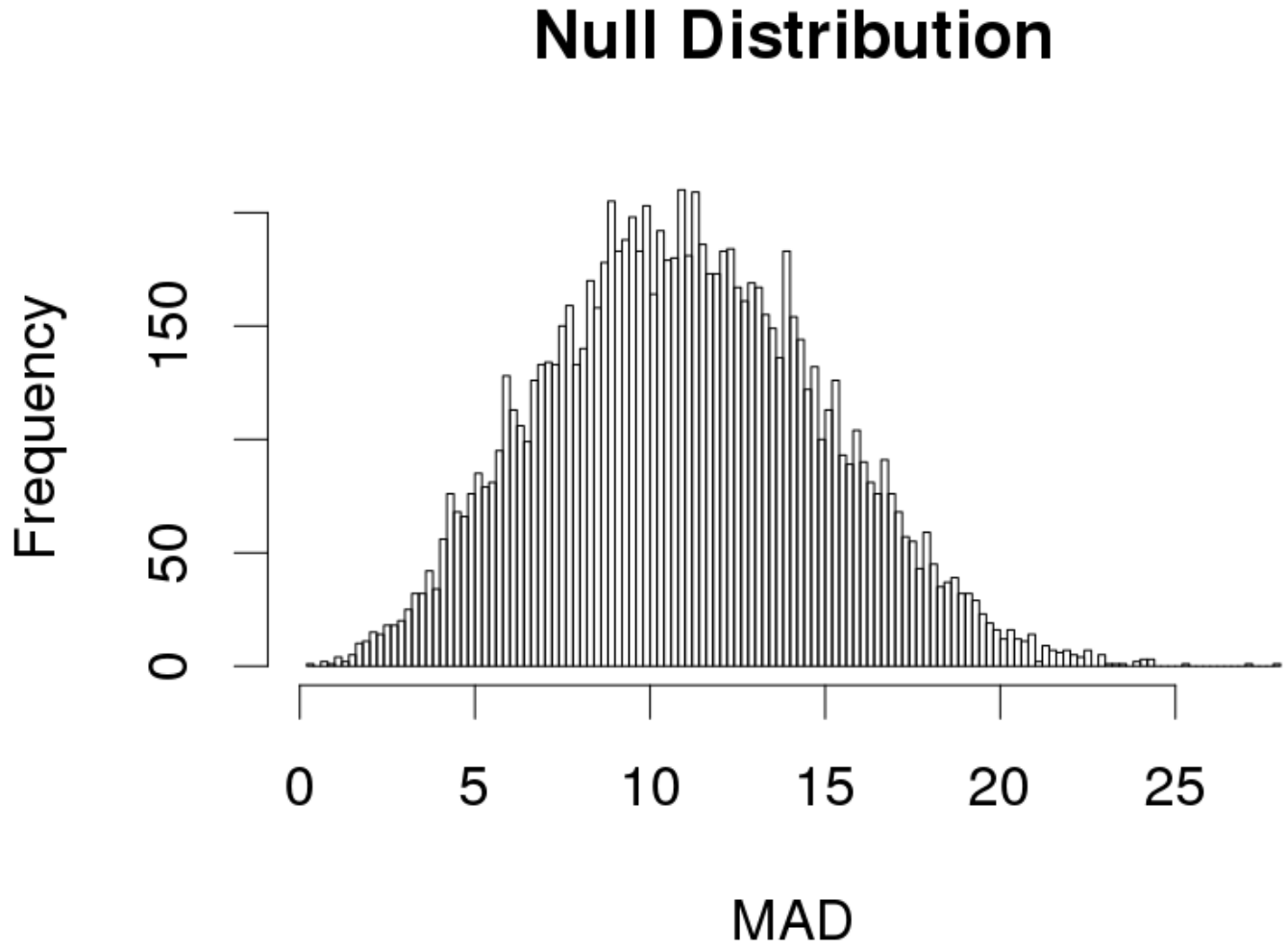
How can we create the null distribution?

3. Create the null distribution!

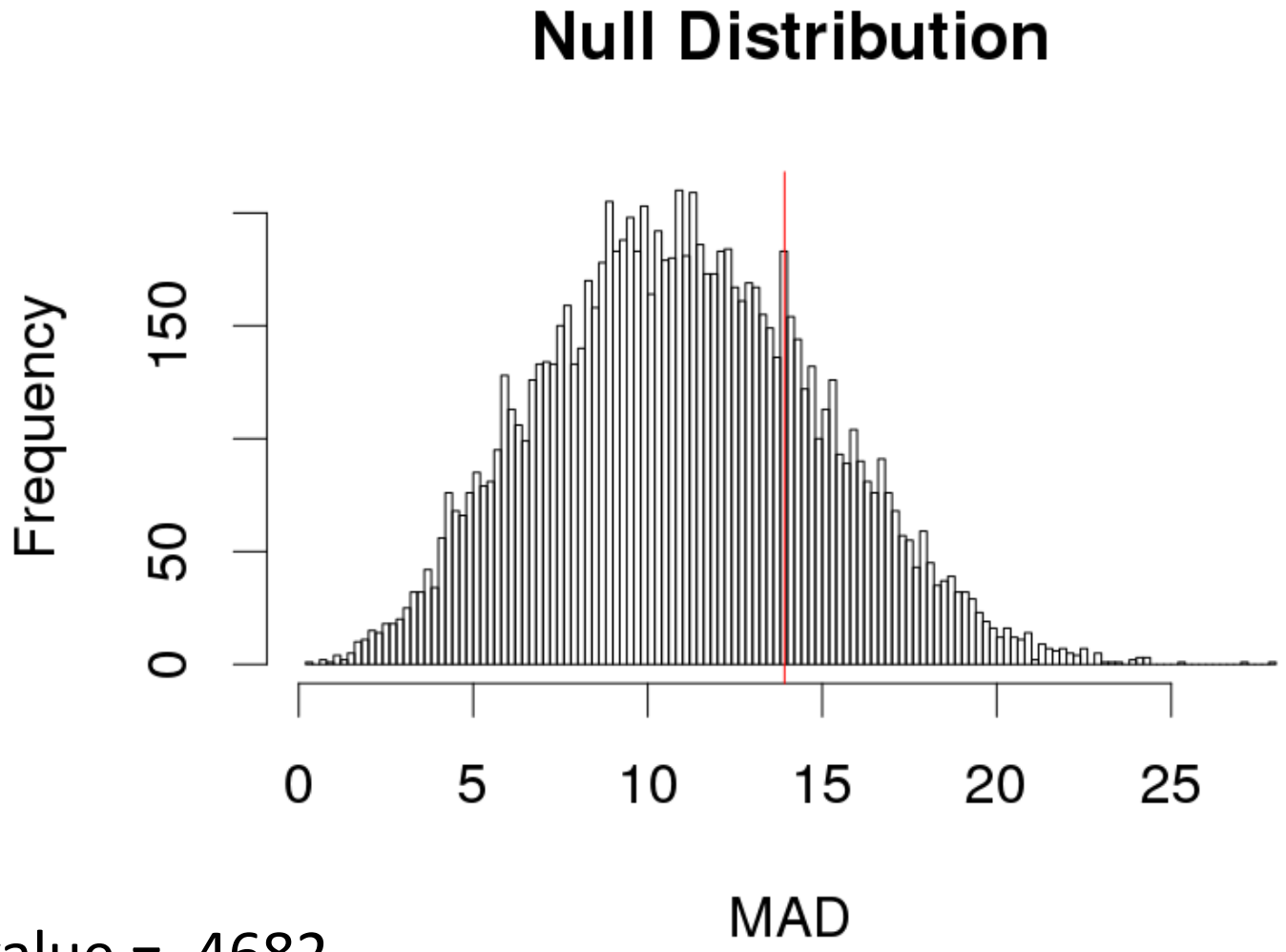


Compute statistics from shuffled groups

Null distribution



P-value



p-value = .4682

Conclusions?

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)

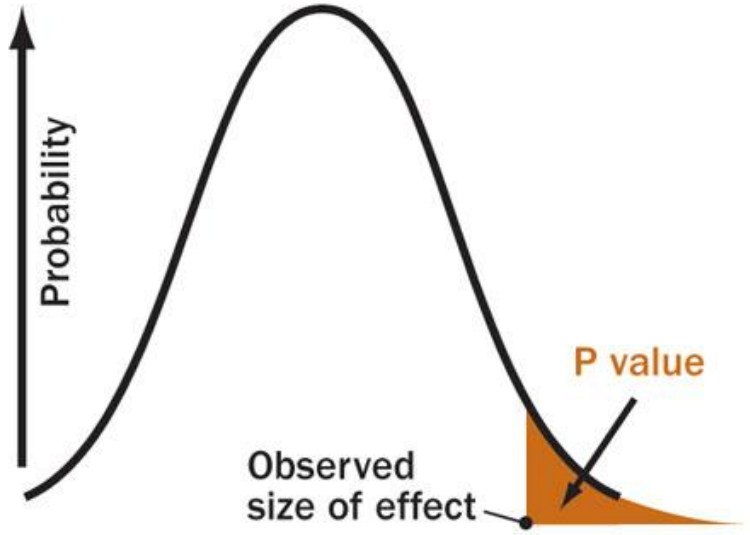
Neyman (1894-1981)

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 so extreme is unlikely when H_0 is true.

Do not reject H_0: if the statistic is not too extreme when H_0 is true. This means the test is inconclusive.

Significance level

The **significance level**, α, for a test of hypothesis is a boundary below which we conclude that a p-value shows statistically significant evidence against the null hypothesis.

The significance level is chosen prior to analyzing the data

- Typical levels: 0.05, 0.01

Formal Statistical Decision Based on a Significance Level

Given a significance level α and a p-value from a sample, we:

Reject H_0 if the p-value is $< \alpha$

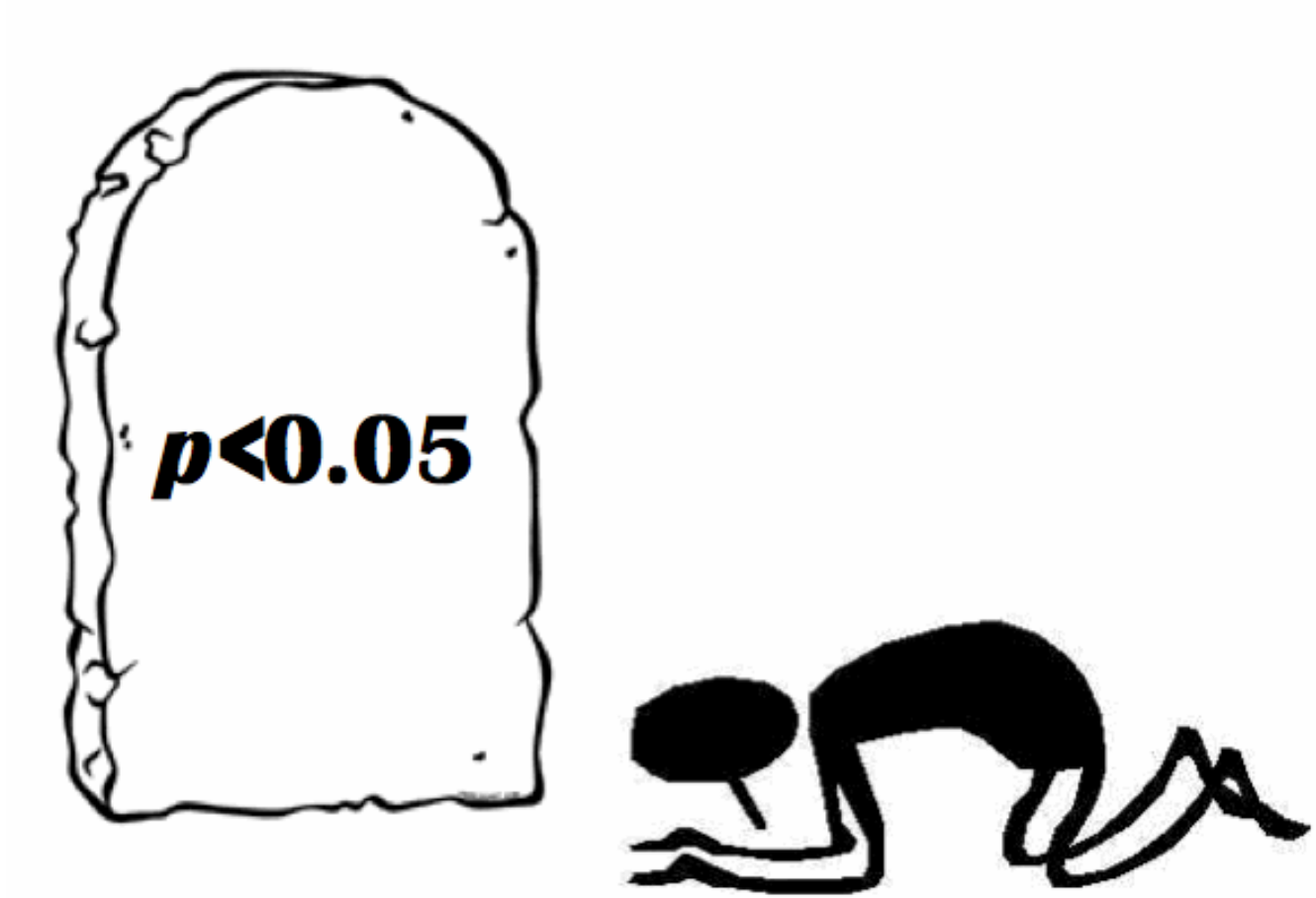
Do not reject H_0 if the p-value is $\geq \alpha$

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 $\alpha = 0.05$)

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


$p<0.05$