Reshaping data,
hypothesis tests for more than 2 means

Overview

Reshaping data

Hypothesis tests for more than 2 means

Reshaping data

Wide vs. Long data

Plotting data using ggplot requires that data is in the right format

- i.e., requires data transformations

Often this involves converting data from a **wide format** to **long format**

Wide data

Person	Age	Height
Bob	32	72
Alice	24	65
Steve	64	70

Narrow data

Person	name	value
Bob	Age	32
Bob	Height	72
Alice	Age	24
Alice	Height	65
Steve	Age	64
Steve	Height	70

```
library(tidyr)
```

tidyr::pivot_longer()

pivot_longer(df, cols) converts data from **wide** to **long**

- Argument **cols**: a vector of strings listing columns to convert to long format
- Converts the data frame into two columns: name and value
  - Column names become categorical variable levels of a new variable called **name**
  - The data in rows become entries in a variable called **value**

Wide data

Person	Age	Height
Bob	32	72
Alice	24	65
Steve	64	70

Long data

Person	name	value
Bob	Age	32
Bob	Height	72
Alice	Age	24
Alice	Height	65
Steve	Age	64
Steve	Height	70

```
pivot_longer(df_wide, cols = c("Age", "Height")
```

tidyr::pivot_wider()

pivot_wider(df, names_from, values_from) converts data from narrow to wide

- Turns the levels of categorical data into columns in a data frame

Narrow data

person	name	value
Bob	Age	32
Bob	Height	72
Alice	Age	24
Alice	Height	65
Steve	Age	64
Steve	Height	70

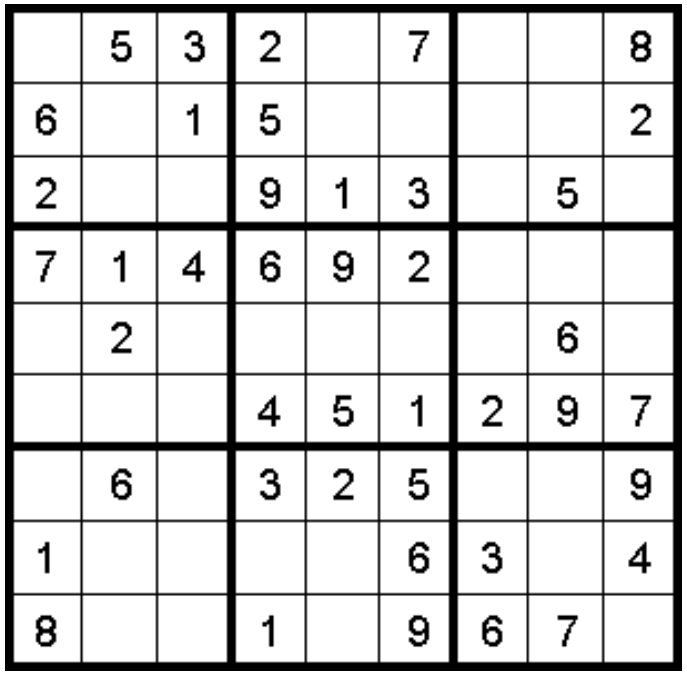
Wide data

Person	Age	Height
Bob	32	72
Alice	24	65
Steve	64	70

pivot_wider(df_long, names_from = name, values_from = value)

Let’s try it in R

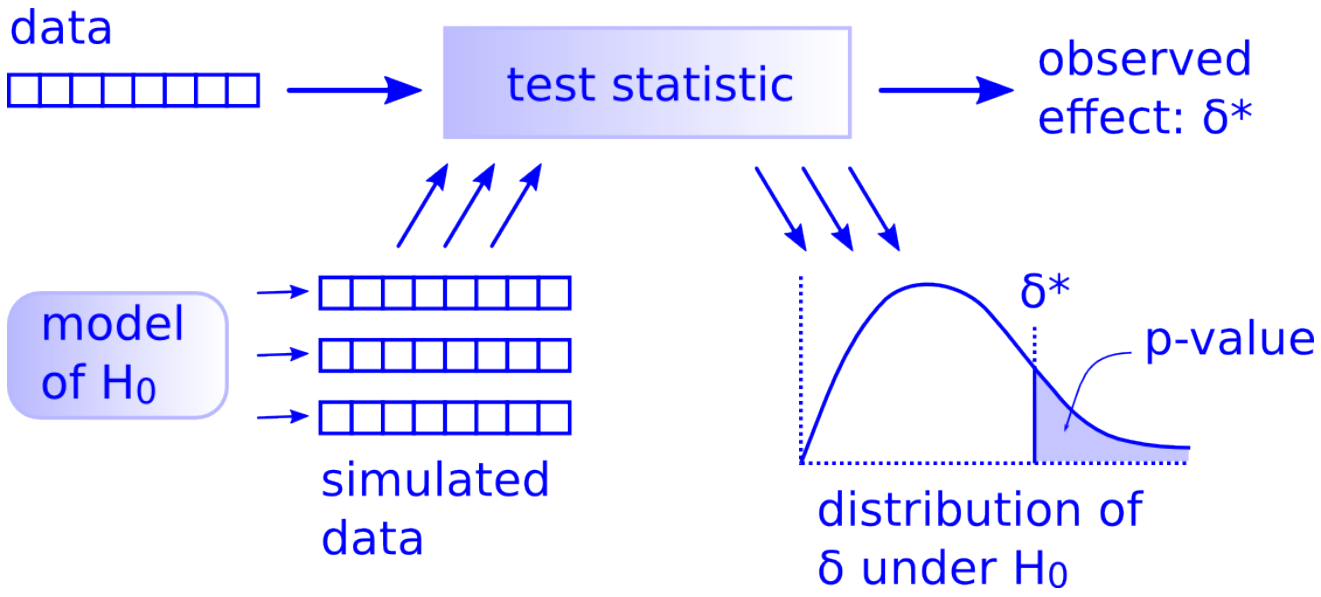
Hypothesis tests for more than two means

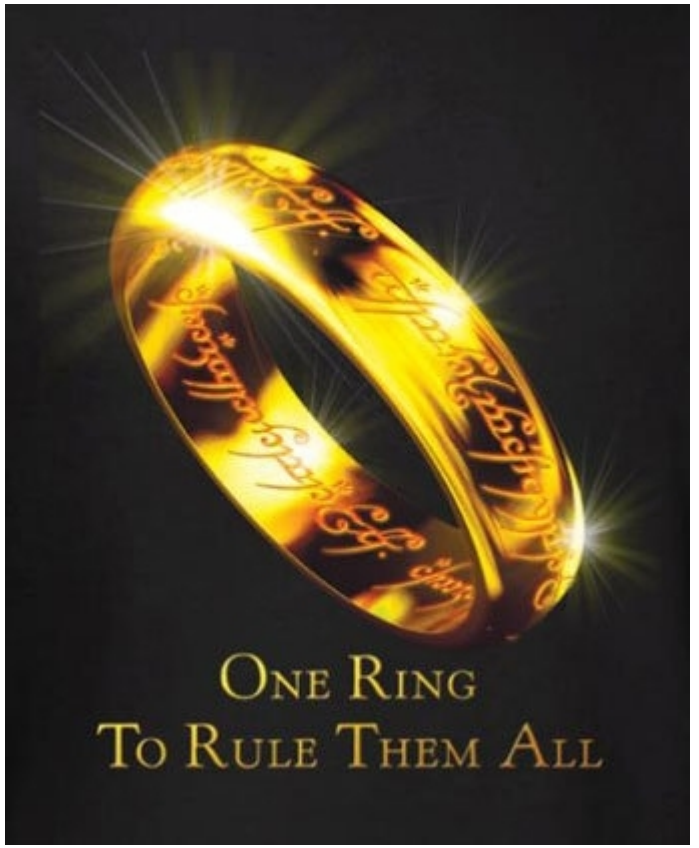


	5	3	2		7			8
6		1	5					2
2			9	1	3		5	
7	1	4	6	9	2			
	2						6	
			4	5	1	2	9	7
	6		3	2	5			9
1					6	3		4
8			1		9	6	7	

Before we start: the big picture…

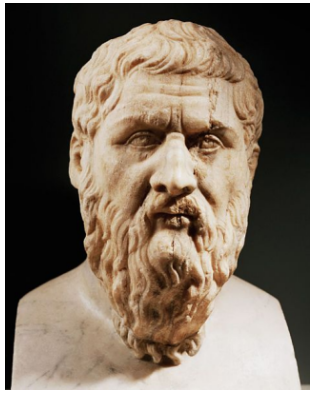
There is only one hypothesis test!





Just follow the 5 hypothesis tests steps!

Five steps of hypothesis testing

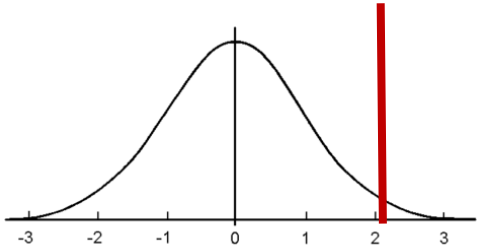


1. State H_0 and H_A
   - Assume Gorgias (H_0) was right
   - α = .05 of the time he will be right, but we will say he is wrong

2. Plot the data and calculate the 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

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)

Sudoku by field

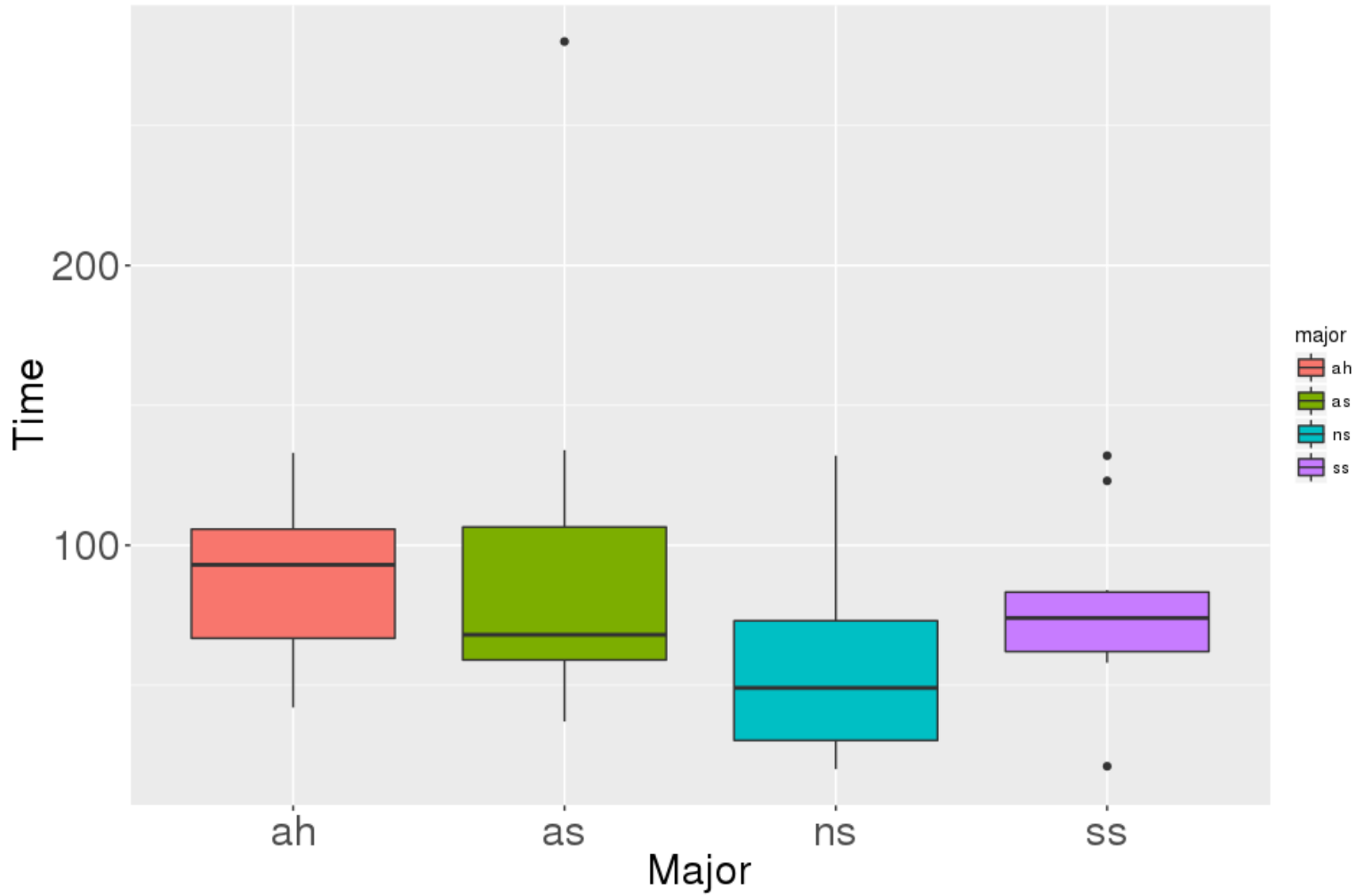
1. State the null and alternative hypotheses!

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

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

What should we do next?

Step 2a: Plot of completion time by major



What should we do next?

Step 2b: Calculating the statistic of interest

What should we use as our statistic?

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

1. Group range statistic:
   $\max \bar{x}_i - \min \bar{x}_i$

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 group range statistic

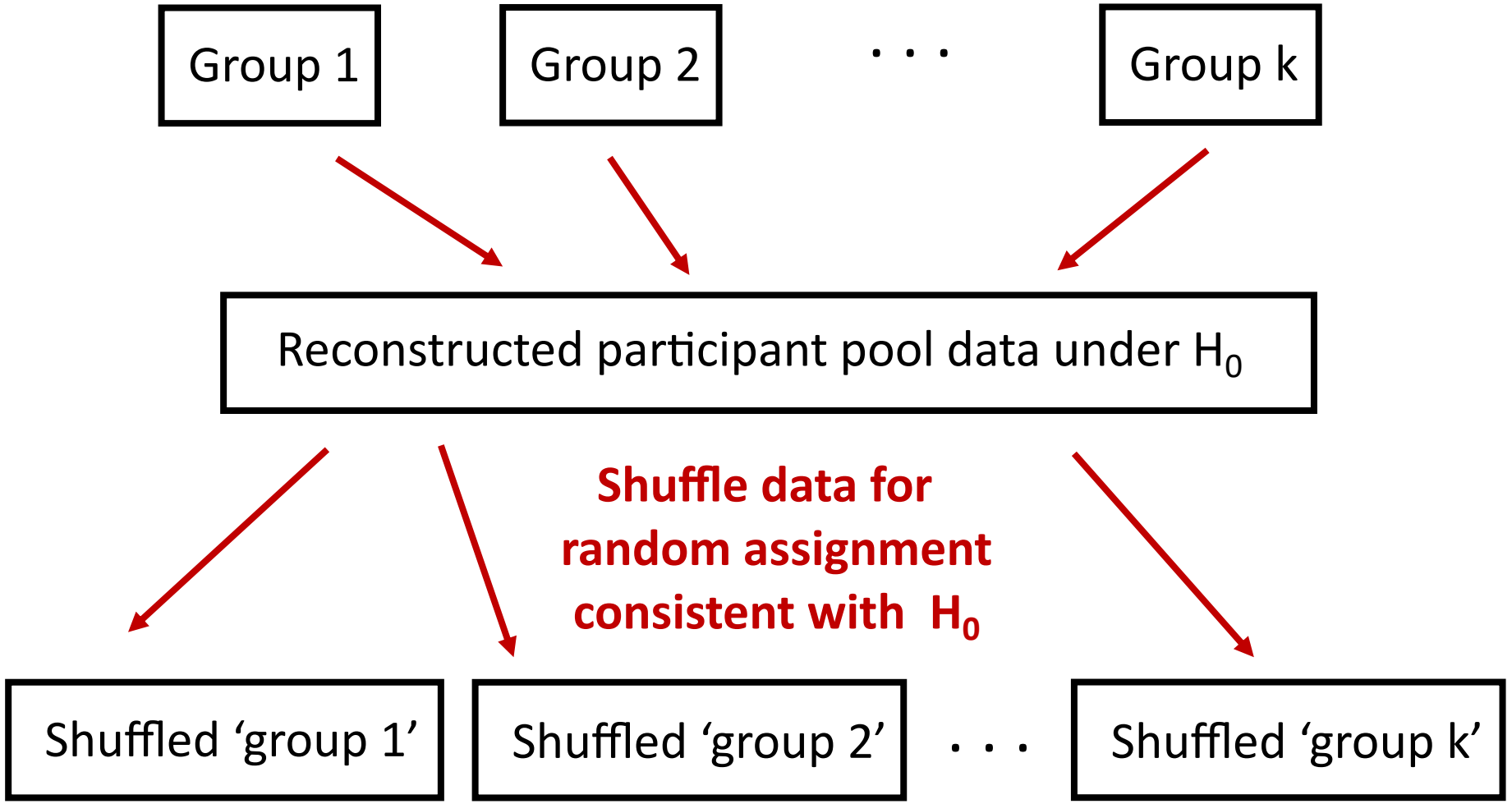
Group range statistic:

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

Observed statistic value = 38.2

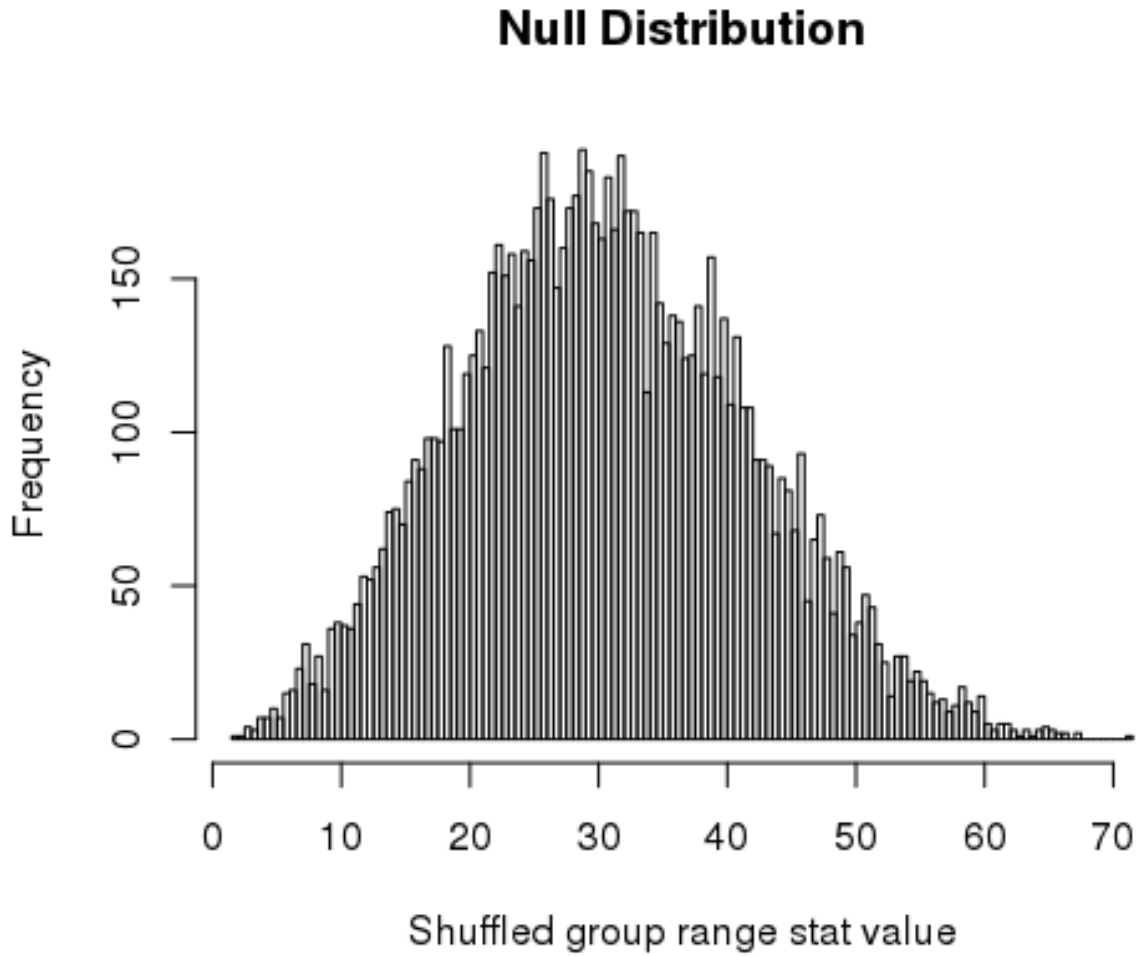
How can we create the null distribution?

Step 3: Create the null distribution!

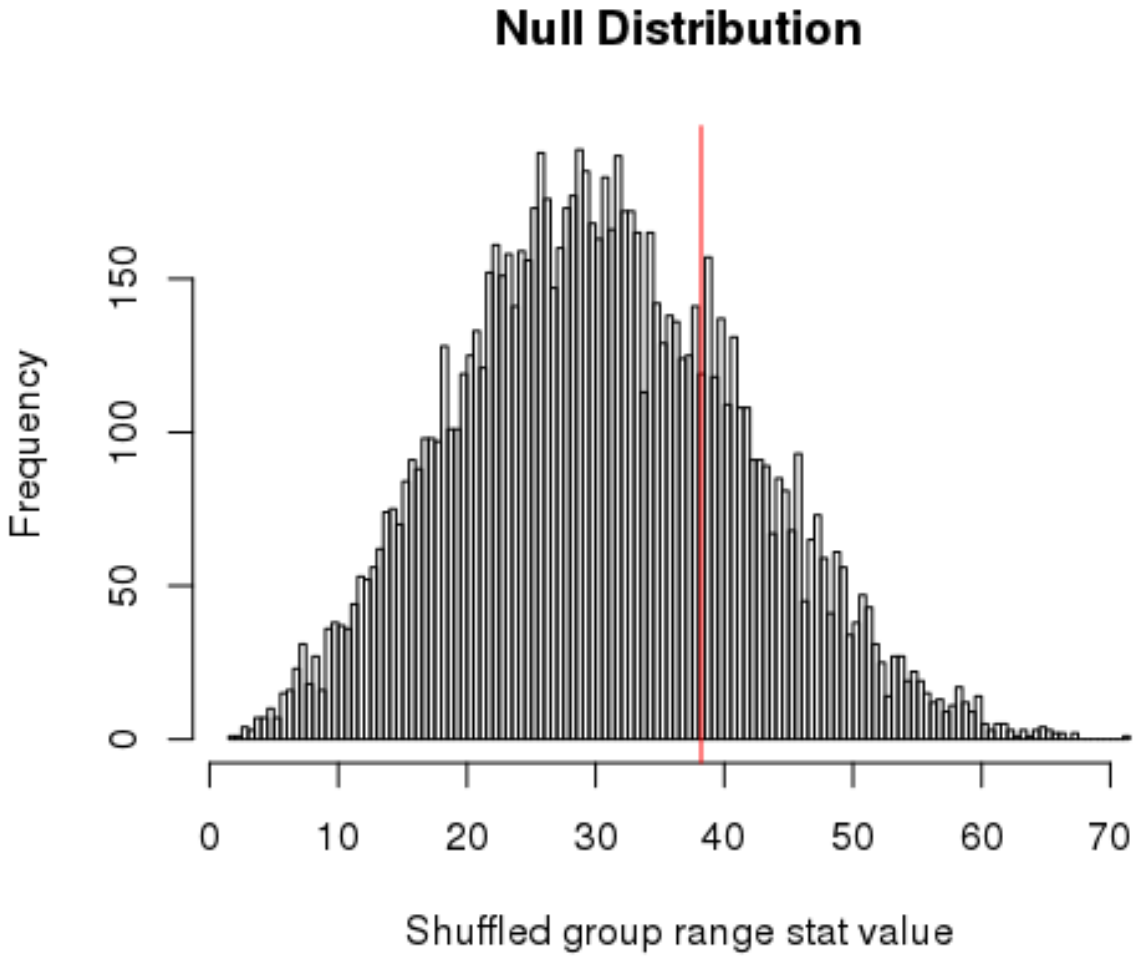


Compute statistics from shuffled groups

Null distribution



Step 4: p-value



p-value = 0.252

Step 5: conclusions?

Let’s try this analysis in R…