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

Review/continuation of hypothesis testing for two means

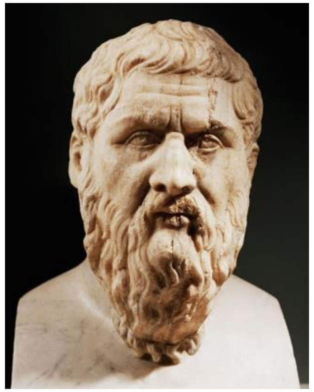
Hypothesis tests for more than two means

If time: theories of hypothesis testing

Hypothesis tests for comparing two means

Question: Can we find out the ***Truth*** of whether the pill effective?

Five steps of hypothesis testing

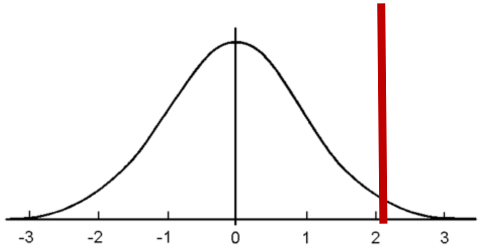


1. State H_0 and H_A
   - Assume Gorgias (H_0) was right

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

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)

1. State the null and alternative hypothesis

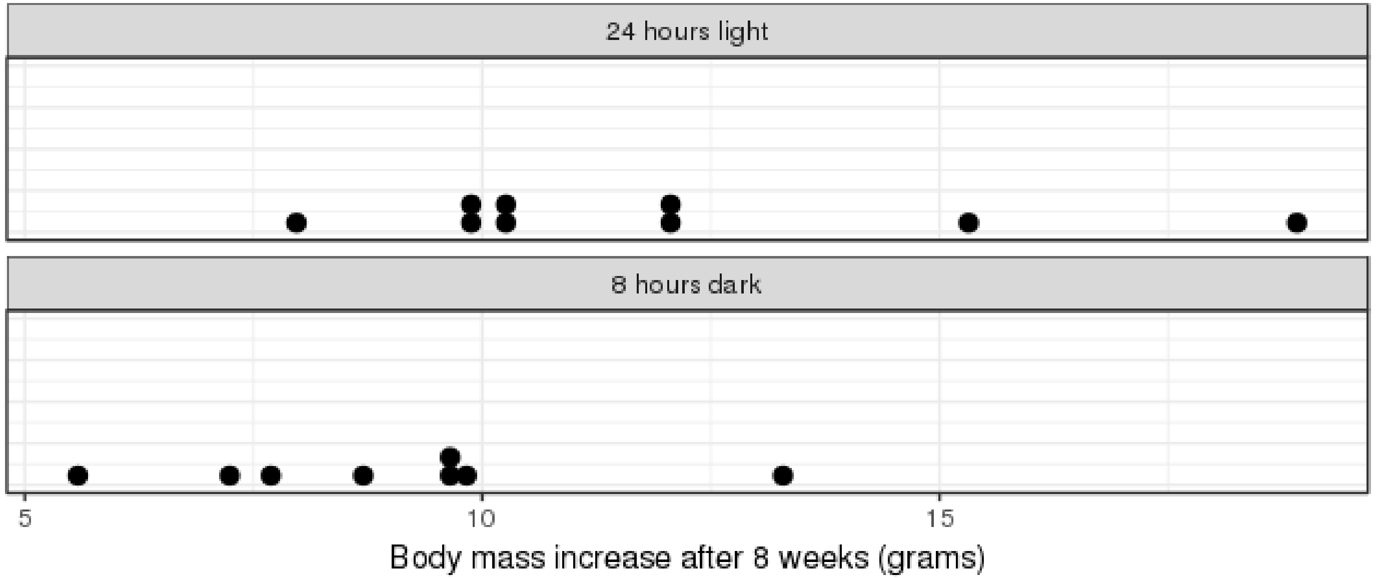
H_0: $\mu_{Light} = \mu_{Dark}$ or $\mu_{Light} - \mu_{Dark} = 0$

H_A: $\mu_{Light} > \mu_{Dark}$ or $\mu_{Light} - \mu_{Dark} > 0$

Hypothesis tests for differences in two group means

What is step 2?

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



Do mice who eat late at night get fat?

You can get the data from:

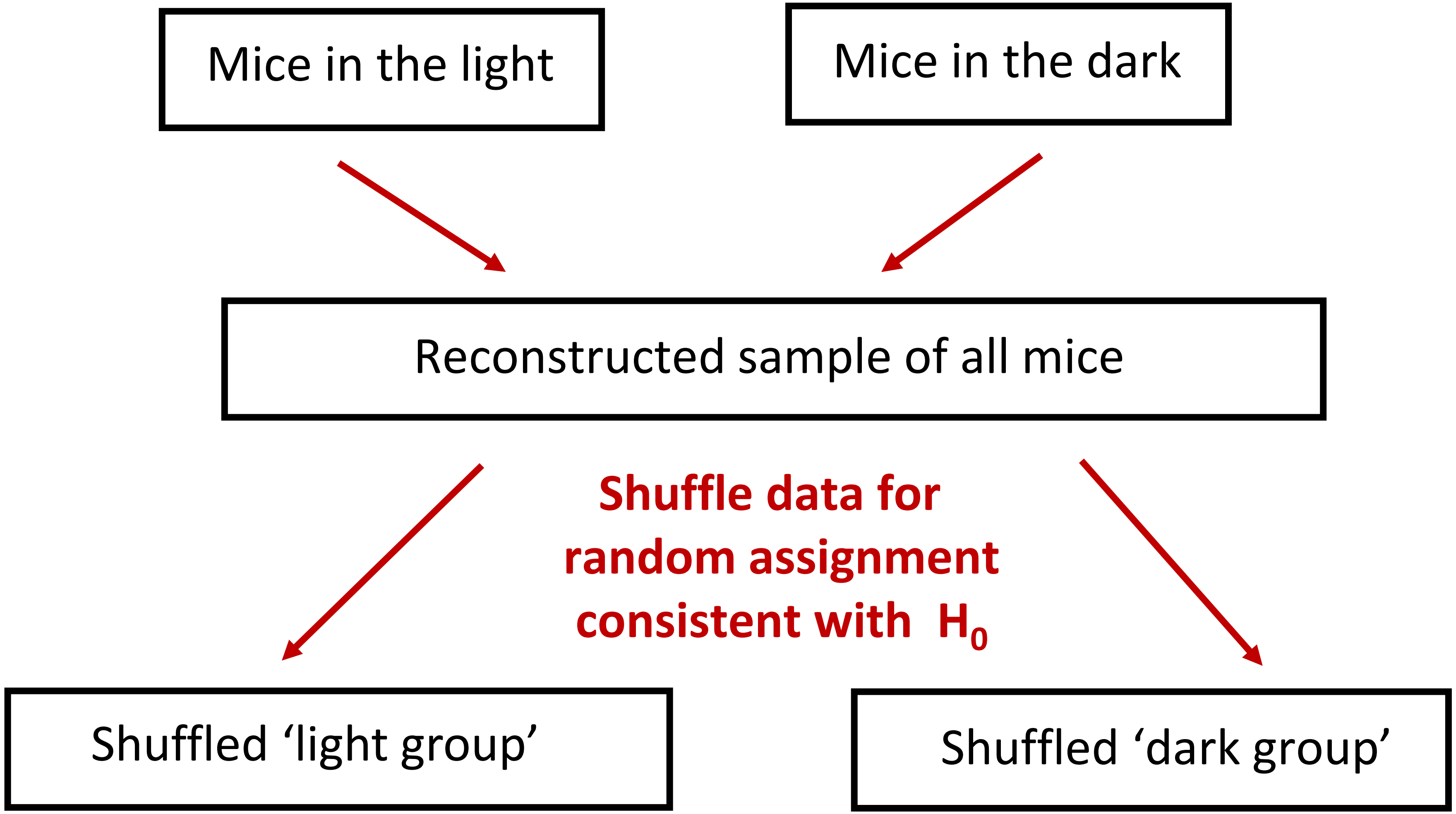
```
download_class_data("mice.Rda")
load("mice.Rda")

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: $\bar{x}_{Shuff\_Light} - \bar{x}_{Shuff\_Dark}$

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  <- shuffle(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)
```

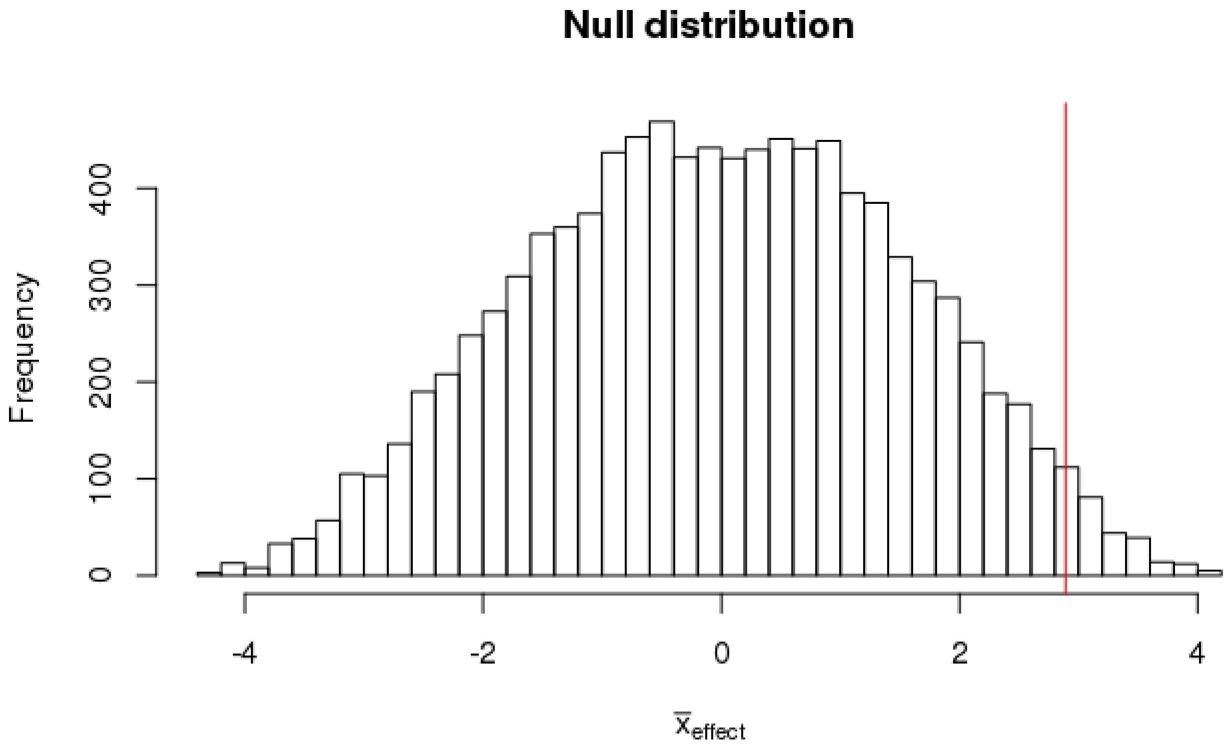
Do mice who eat late at night get fat?

How do we create a full null distribution?

```
null_dist <- do_it(10000)  * {

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

}
```

Do mice who eat late at night get fat?

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



What do we do next?

Do mice who eat late at night get fat?

Get the p-value

```
p_val <- pnull(obs_stat, null_dist, lower.tail = FALSE)
```

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 step of hypothesis testing?

Sudoku by field

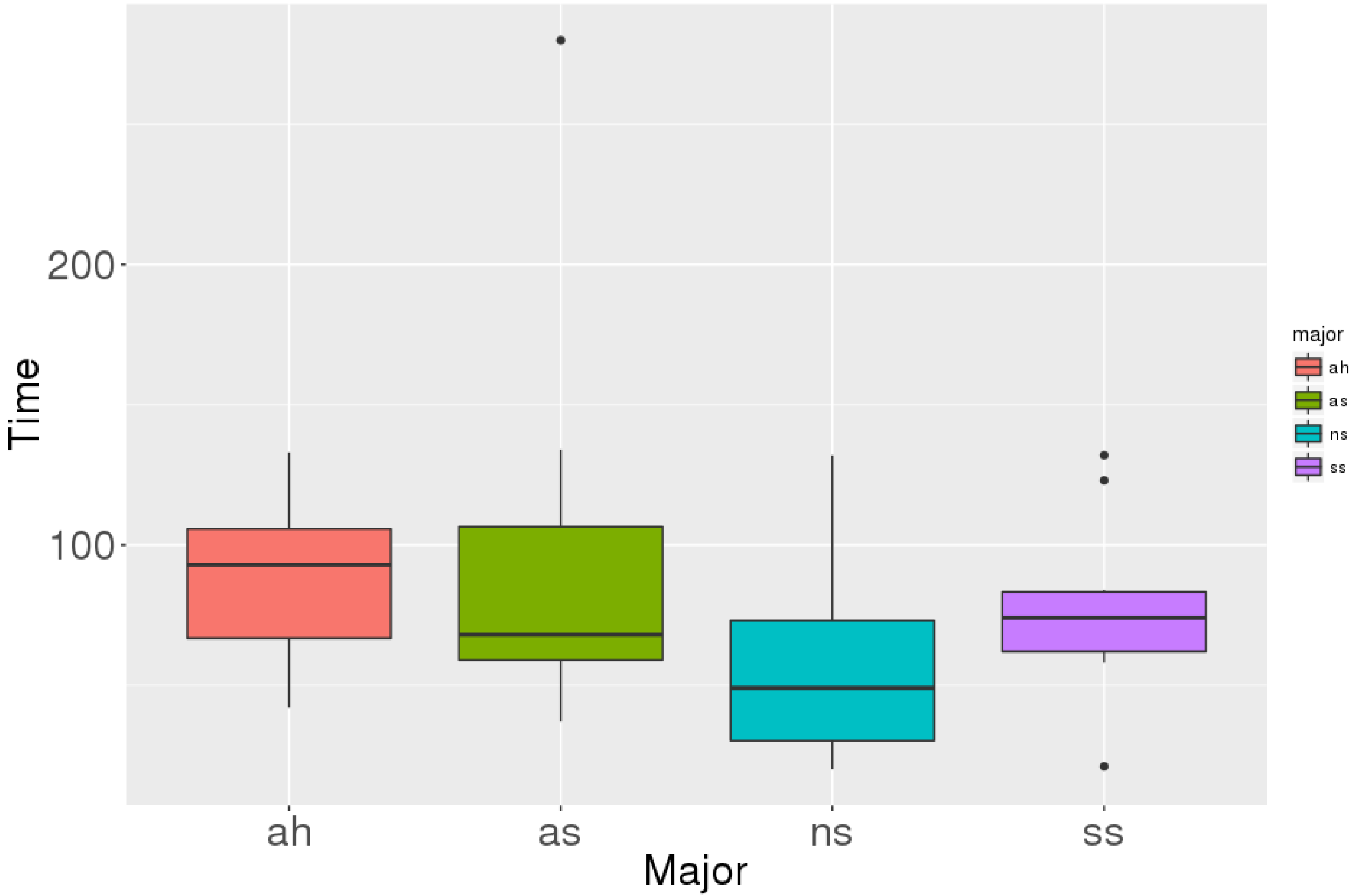
1. State the null and alternative hypotheses!

$\mathbf{H_0}$: $\mu_{as} = \mu_{ns} = \mu_{ss} = \mu_{ah}$
$\mathbf{H_A}$: $\mu_i \neq \mu_j$ for one pair of fields of study

What should we do next?

Let's plot the data first…

Step 0: Plot of completion time by major



What should we do next?

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

Thoughts on the statistic of interest?

Comparing multiple means

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

1. <u>Group range statistic:</u>

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

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

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

3. <u>F statistic:</u>

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}$$

Using the MAD statistic

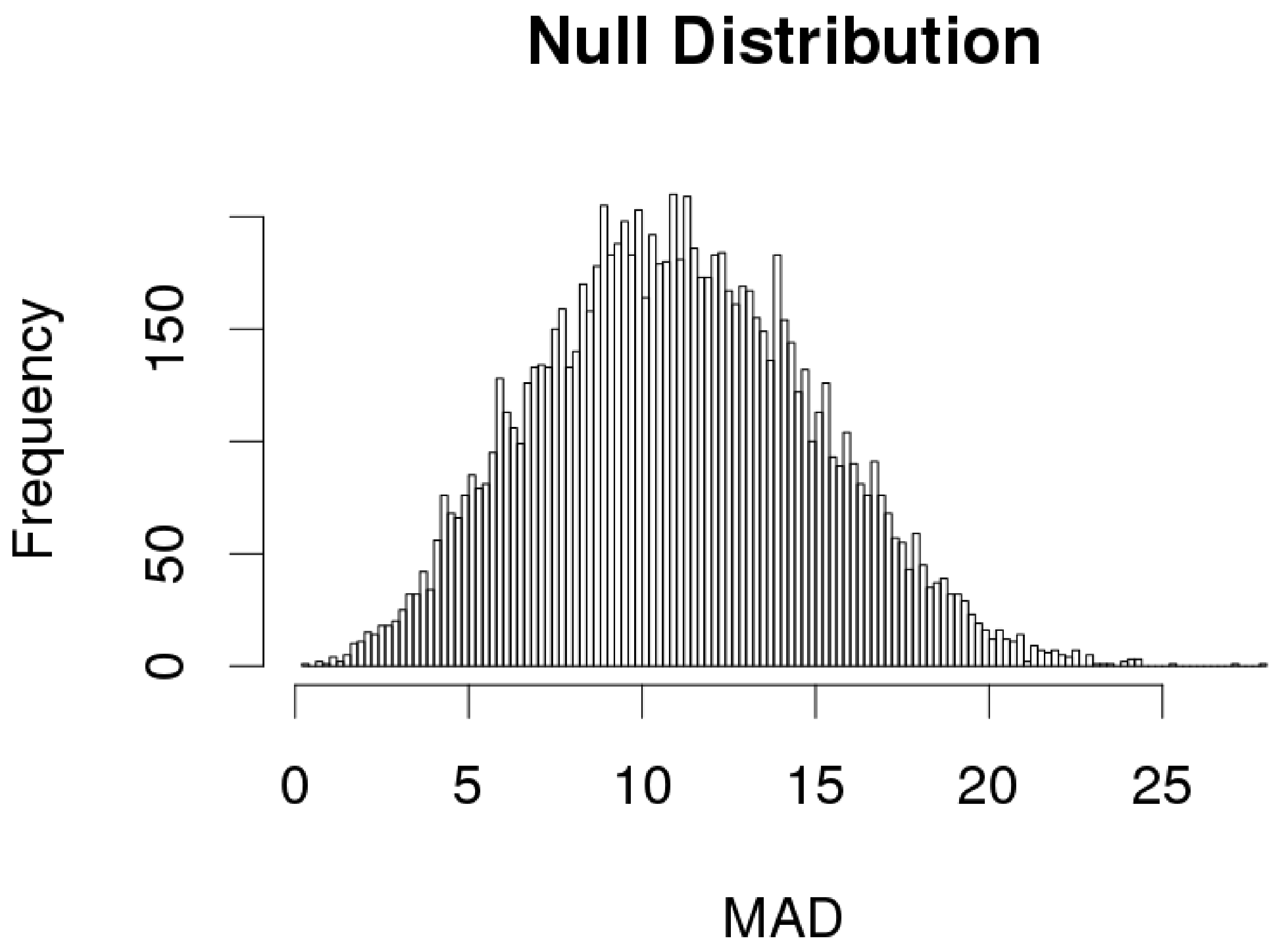
Mean absolute difference (MAD):

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

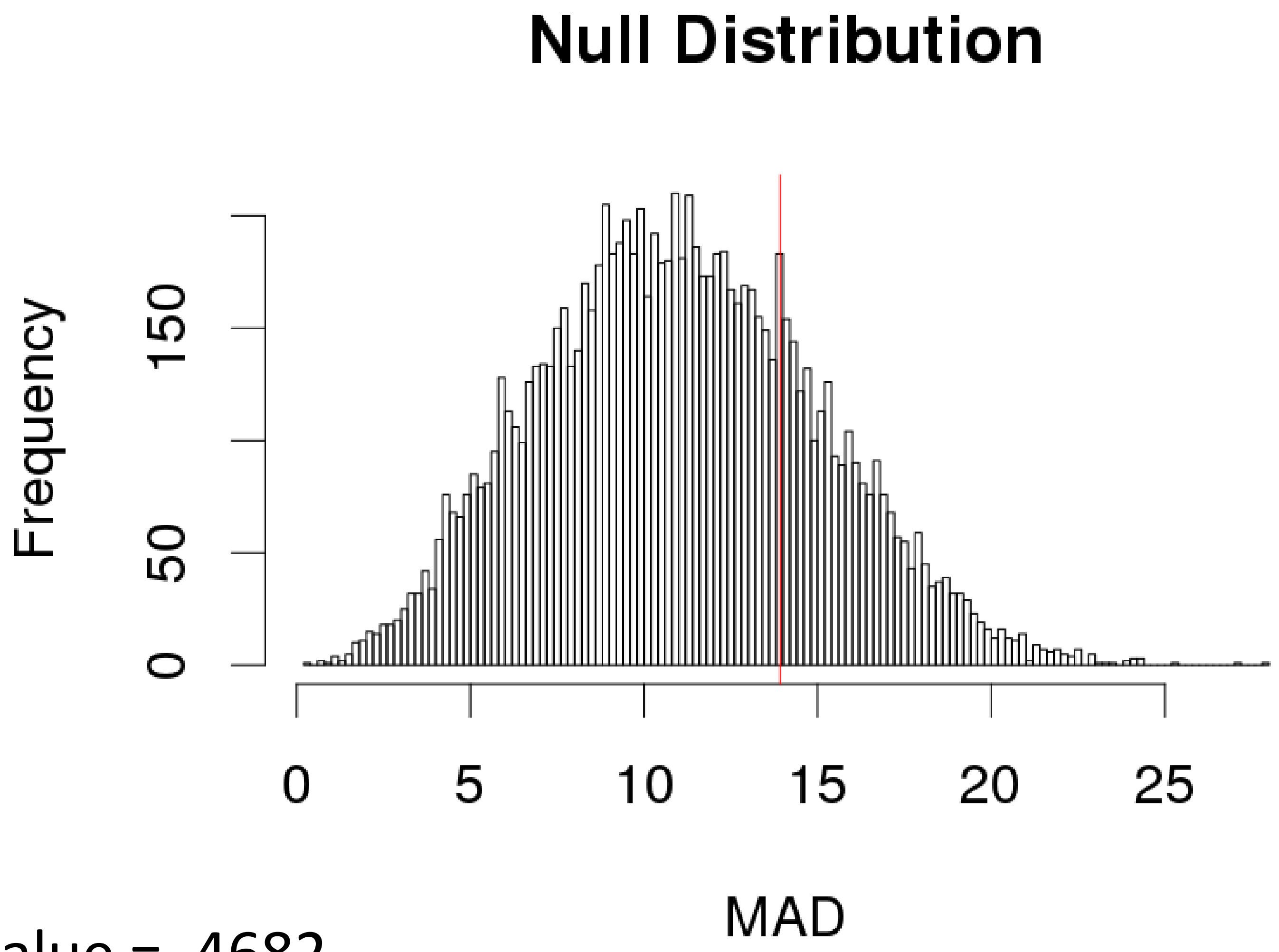
Observed statistic value = 13.92

How can we create the null distribution?

Null distribution



P-value



p-value = .4682

Conclusions?


this sudoku is really hard!
that's a crossword puzzle
hmmm...
ohhh!

Let’s try this analysis in R…

Either run:

```
reinstall_class_package()
```

Or use workspace 3

- Link is on Canvas

```
# get the data
library(ClassTools)
download_class_data("MajorPuzzle.txt")
sudoku_data <- read.table("MajorPuzzle.txt", header = TRUE)
```

Let’s try this analysis in R…

```r
# get the data
library(ClassTools)
download_class_data("MajorPuzzle.txt")
Sudoku_data <- read.table("MajorPuzzle.txt", header = TRUE)

# Extract vectors from the data frame  (how do we do this?)
completion_time <- sudoku_data$time
major <- sudoku_data$major
```

Let's try this analysis in R…

We can get the MAD statistic using the get_MAD_stat() function

get_MAD_stat(data_vector, grouping_vector)

- data_vector: a vector of quantitative data
- grouping_vector: a vector indicating which group the quantitative data is in

Can you get the MAD statistic for the sudoku data?

```
obs_stat <-  get_MAD_stat(completion_time, major)
```

Can you visualize the data?

```
boxplot(completion_time ~ shuffled_majors)
```

Let’s try this analysis in R…

Q: How could we create one point in a null distribution?

- A: Shuffle the grouping_vector (major vector) and calculate the MAD statistic

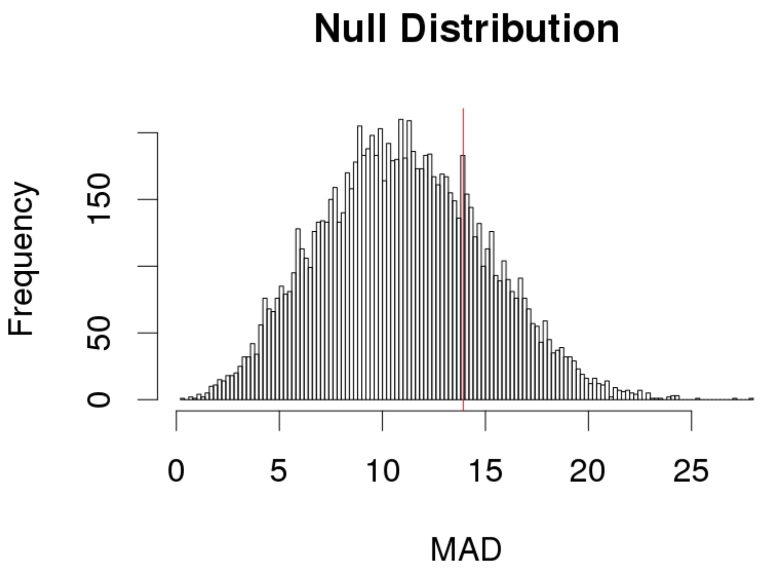
Q: How can we do this in R?

```
shuffled_majors <- shuffle(major)

get_MAD_stat(completion_time, shuffled_majors)
```

Let’s try this analysis in R…

Q: How can we create a full null distribution?

```
null_dist <- do_it(10000) * {
    shuffled_majors <- shuffle(major)
    get_MAD_stat(completion_time, shuffled_majors)
}

# visualize the null distribution
hist(null_dist, breaks = 100)
abline(v = obs_stat, col = "red")
```



Let’s try this analysis in R…

Q: What do we do next and how do we do it?

- A: We get the p-value

```
pnull(obs_stat, null_dist, lower.tail = FALSE)
```