CSCI 443: LECTURE 20
ANOVA

Professor David Harrison

OFFICE HOURS

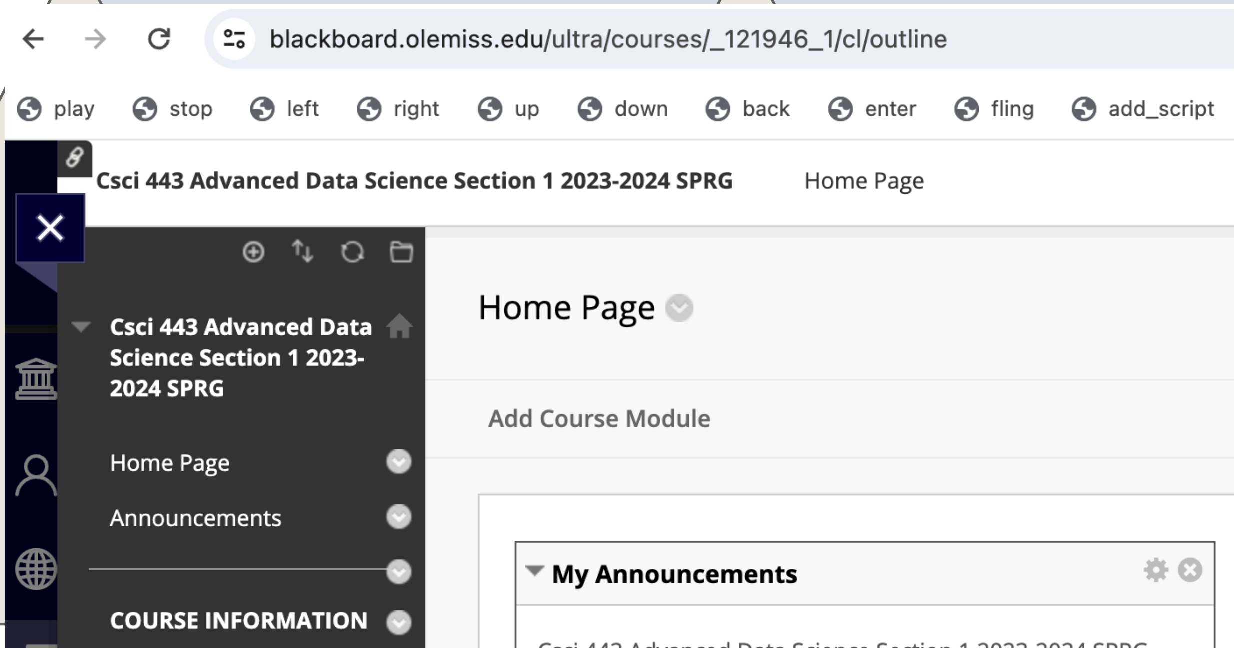
Tuesday	4:00-5:00 PM
Wednesday	12:30-2:30 PM

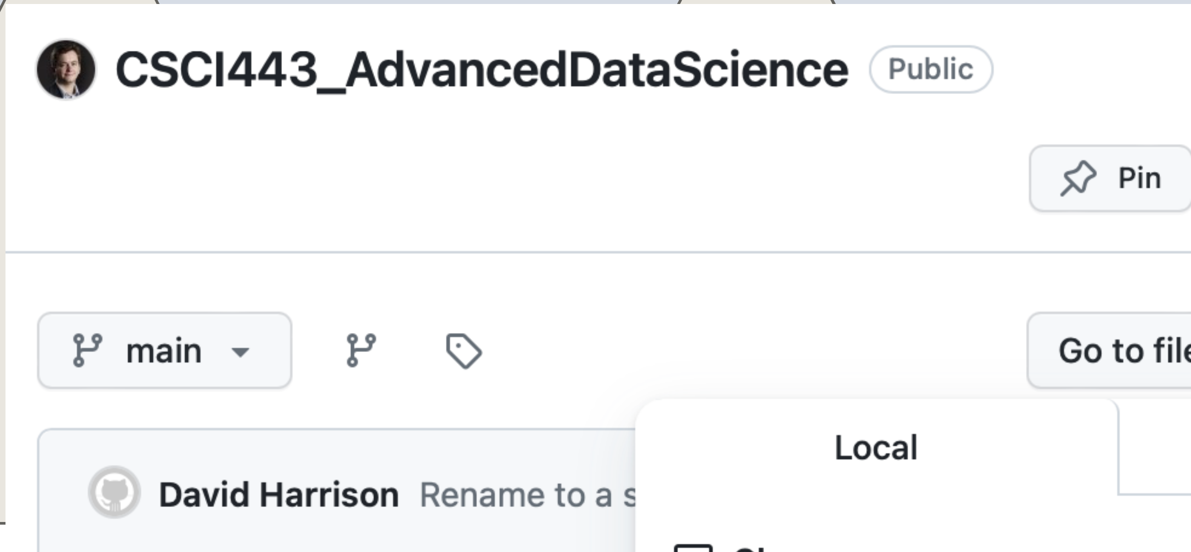
BLACKBOARD & GITHUB

Slides and a jupyter notebook for lecture 19 are on blackboard and in GitHub.

The project is at

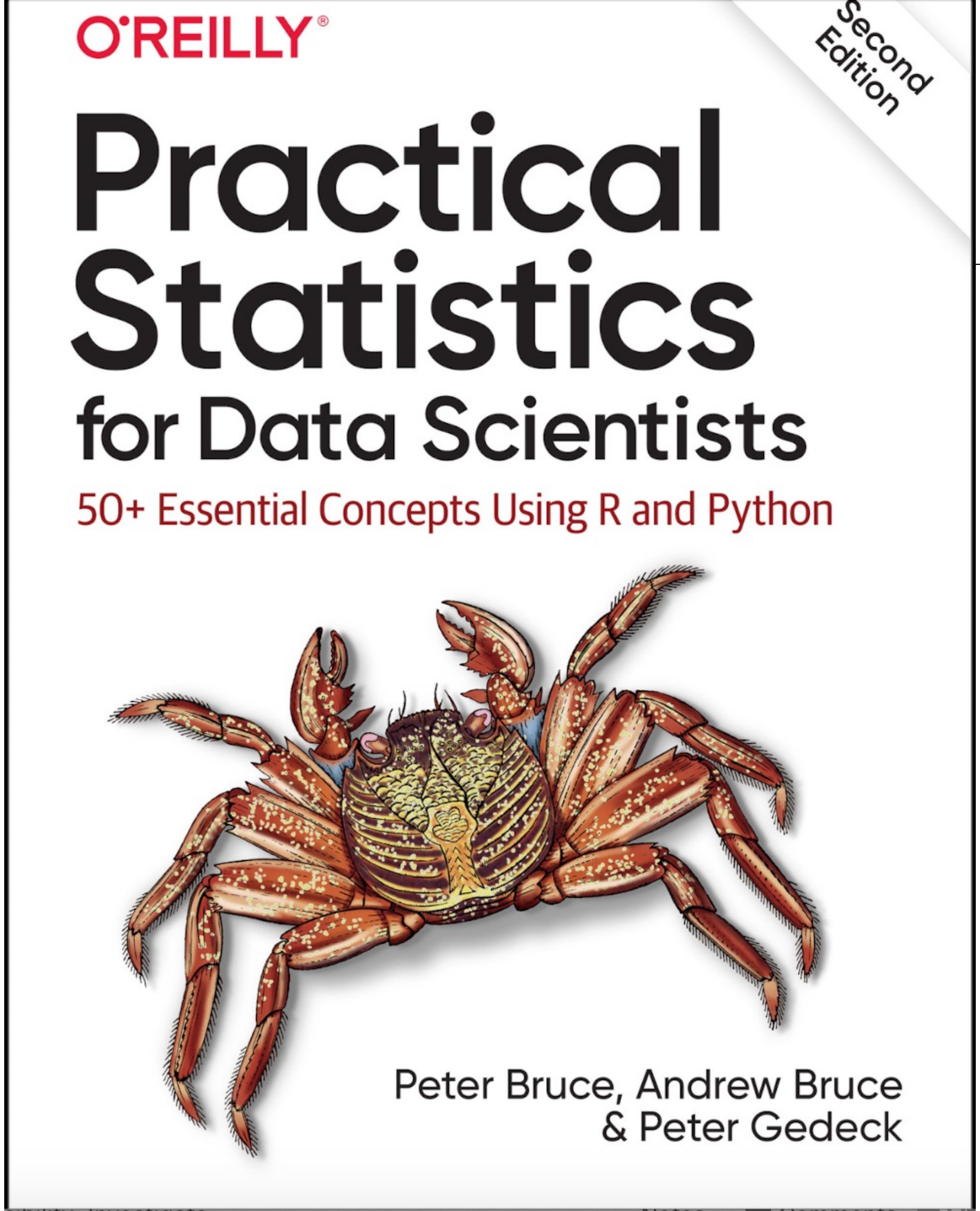
https://github.com/dosirrah/CSCI443_AdvancedDataScience





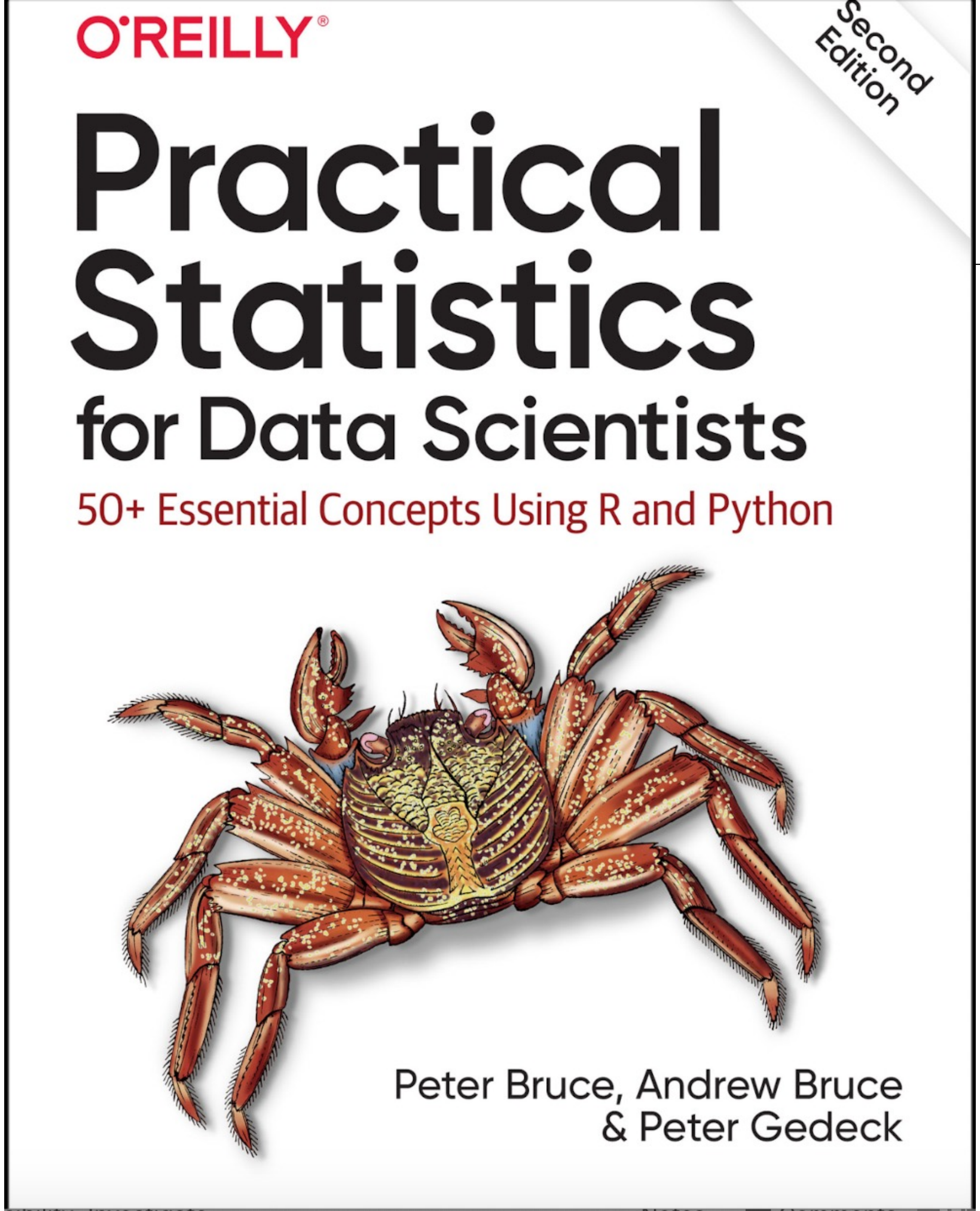
READ ABOUT

- chapter 3: experiments, hypothesis testing
  - ANOVA
  - Chi-square



THINGS I WANT TO COVER TODAY

- ANOVA
- F-Statistic

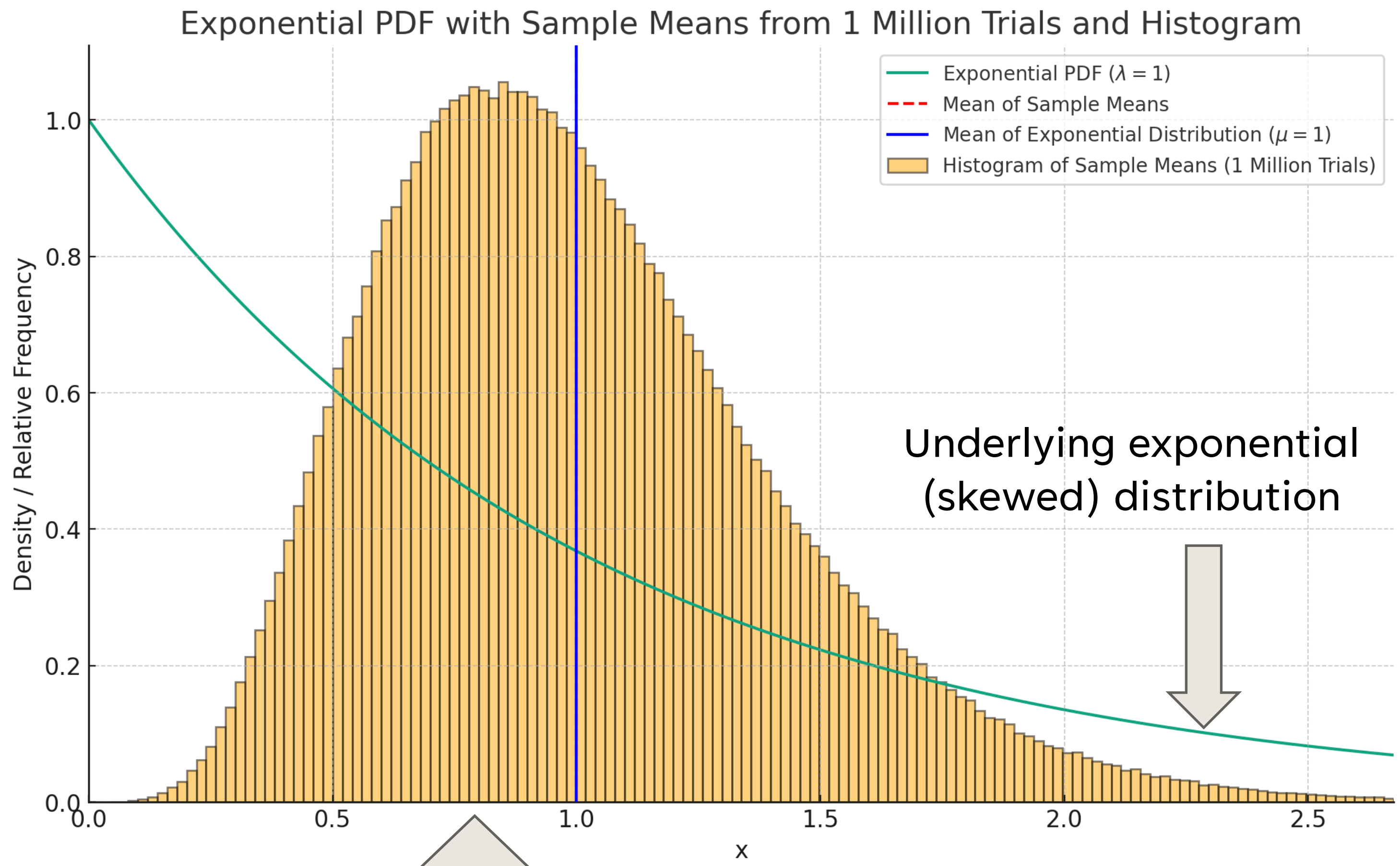


PREVIOUSLY: PERMUTATION TEST

A *permutation test* is an alternative to hypothesis testing using either a t-distribution or Gaussian distribution as an approximation of the sampling distribution.

Used when

1) have small sample sizes
   - **CLT doesn't apply.**
2) the sampling distribution doesn't look normal.
3) uncertain of homoscedasticity (uncertain of equal variances)
4) have complex or uncommon statistical models
5) desire simplicity and robustness



Underlying exponential (skewed) distribution

Non-normal sampling distribution of the mean. Exhibits right skew.

PREVIOUSLY: HOW PERMUTATION TESTS WORK

Used with null hypothesis testing for A/B.

1. Combine samples from different groups into a single data set.
2. Shuffle the combined data set and randomly draw without replacement same size as group A.
3. Draw without replacement same size as group B
4. Measure test statistic.
5. Repeat until R times to build a permutation distribution.

The permutation distribution is an estimate of the sampling distribution.

PREVIOUSLY: HOW PERMUTATION TESTS WORK

We can combine into a single dataset since we are proceeding from the assumption that the null hypothesis is true.

If it is true then the A and B at least have the same population mean.

1. Combine samples from different groups into a single data set.
2. Shuffle the combined data set and randomly draw without replacement same size as group A.
3. Draw without replacement same size as group B
4. Measure test statistic.
5. Repeat until R times to build a permutation distribution.

The permutation distribution is an estimate of the sampling distribution.

PREVIOUSLY: EXAMPLE: INCANDESCENT VS. LED LIGHTS

We have been tasked with confirming that LED lights last longer than incandescent lights.

We gathered data from 100 light bulbs of each kind under identical simulated use patterns.

We continued the trial until 30% of the light bulbs fail for each kind.

We therefore have 30 failures of each kind in our sample sets.

PREVIOUSLY: EXAMPLE: INCANDESCENT VS. LED LIGHTS

let X = lifespan of an incadescent light bulb (in years)

let Y = lifespan of an led light bulb (in years)

let $H_0 : \mu_x = \mu_y$

let $H_A : \mu_x \neq \mu_y$

PREVIOUSLY: EXAMPLE: INCANDESCENT VS. LED LIGHTS

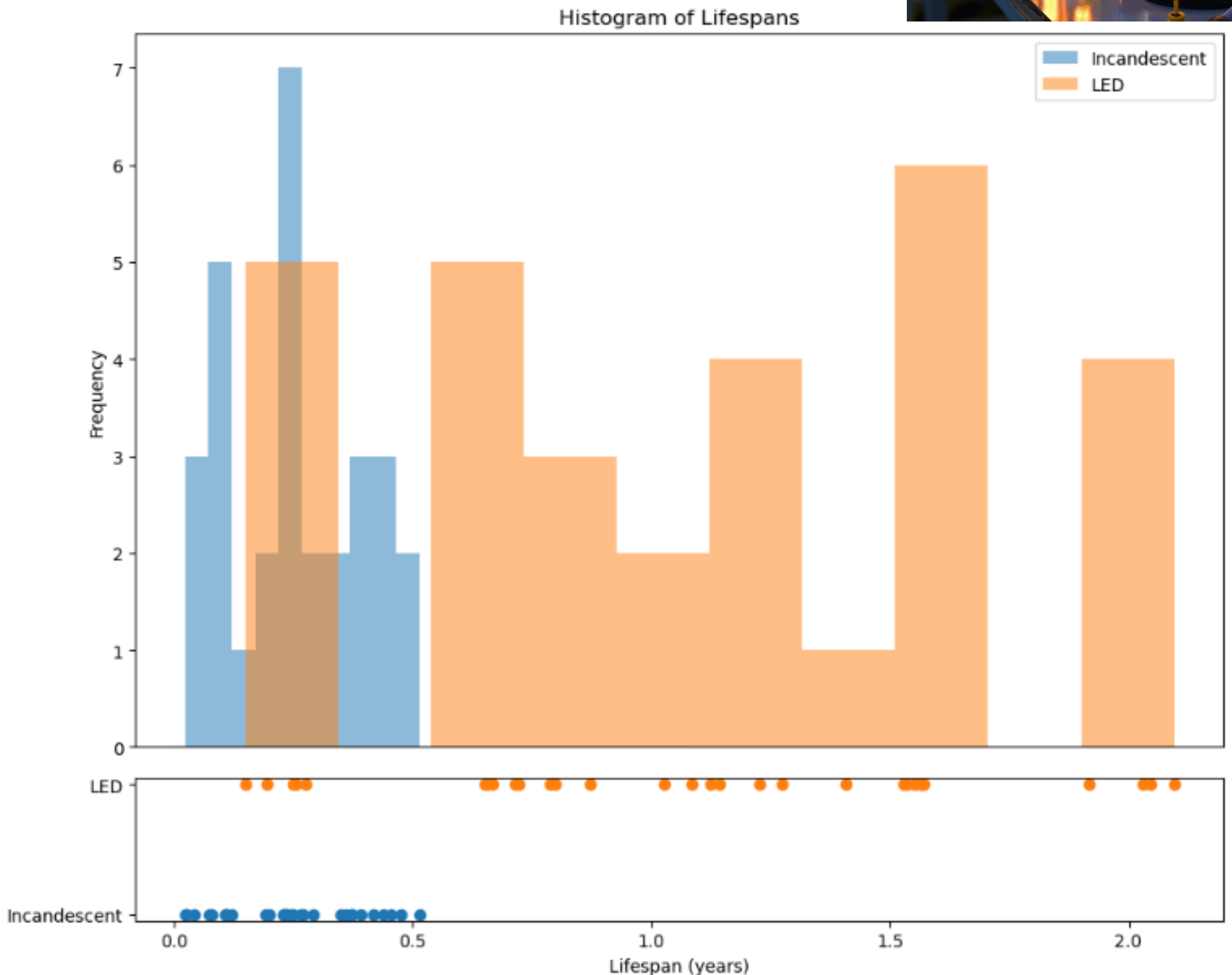
Neither distribution looks Gaussian.

- Seems to few samples for CLT to apply.

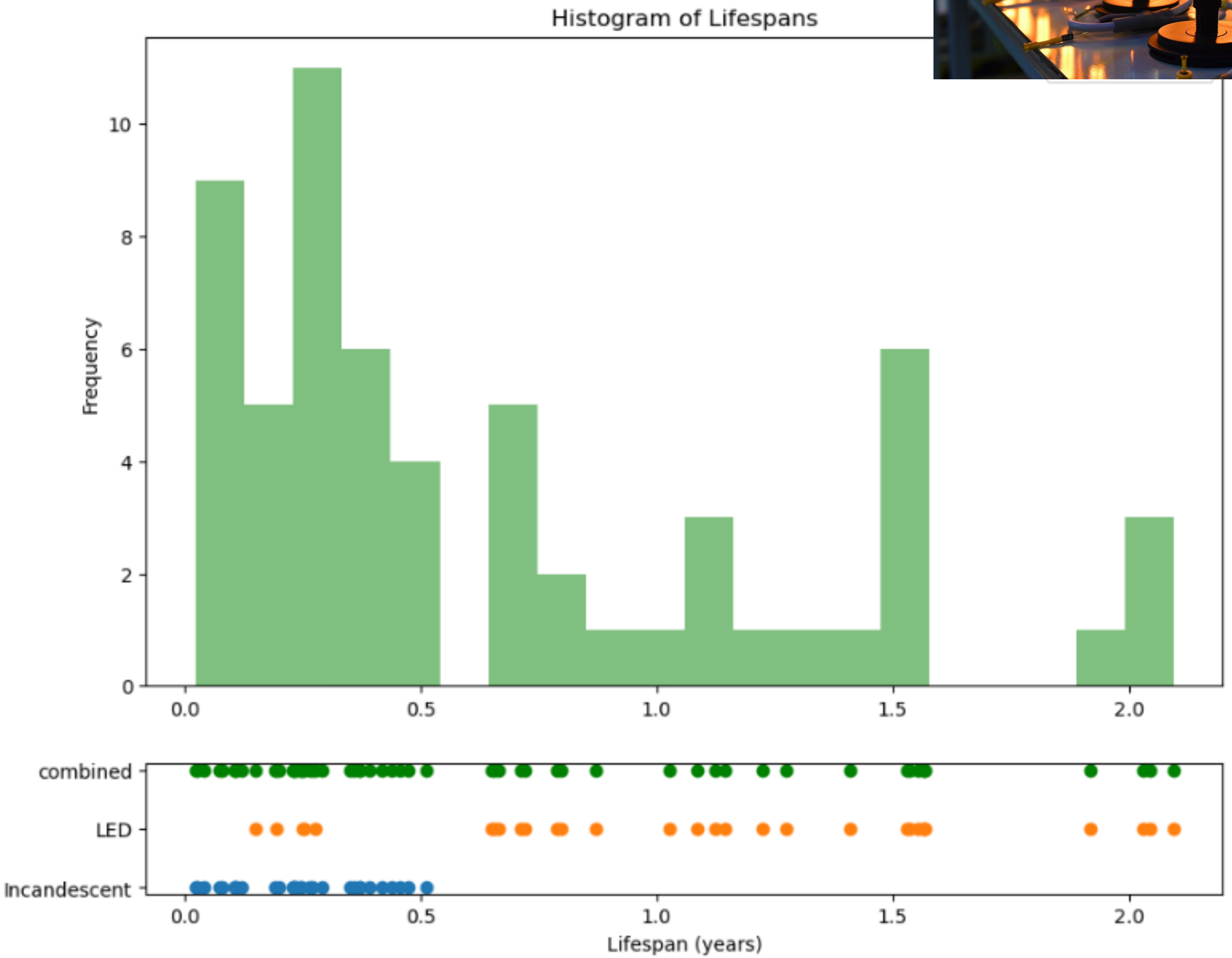
T-distribution is based on a Gaussian assumption

- Used when sample mean and sample variance are computed from the same samples.
- So no t-test.

When this happens, permutation tests make sense.

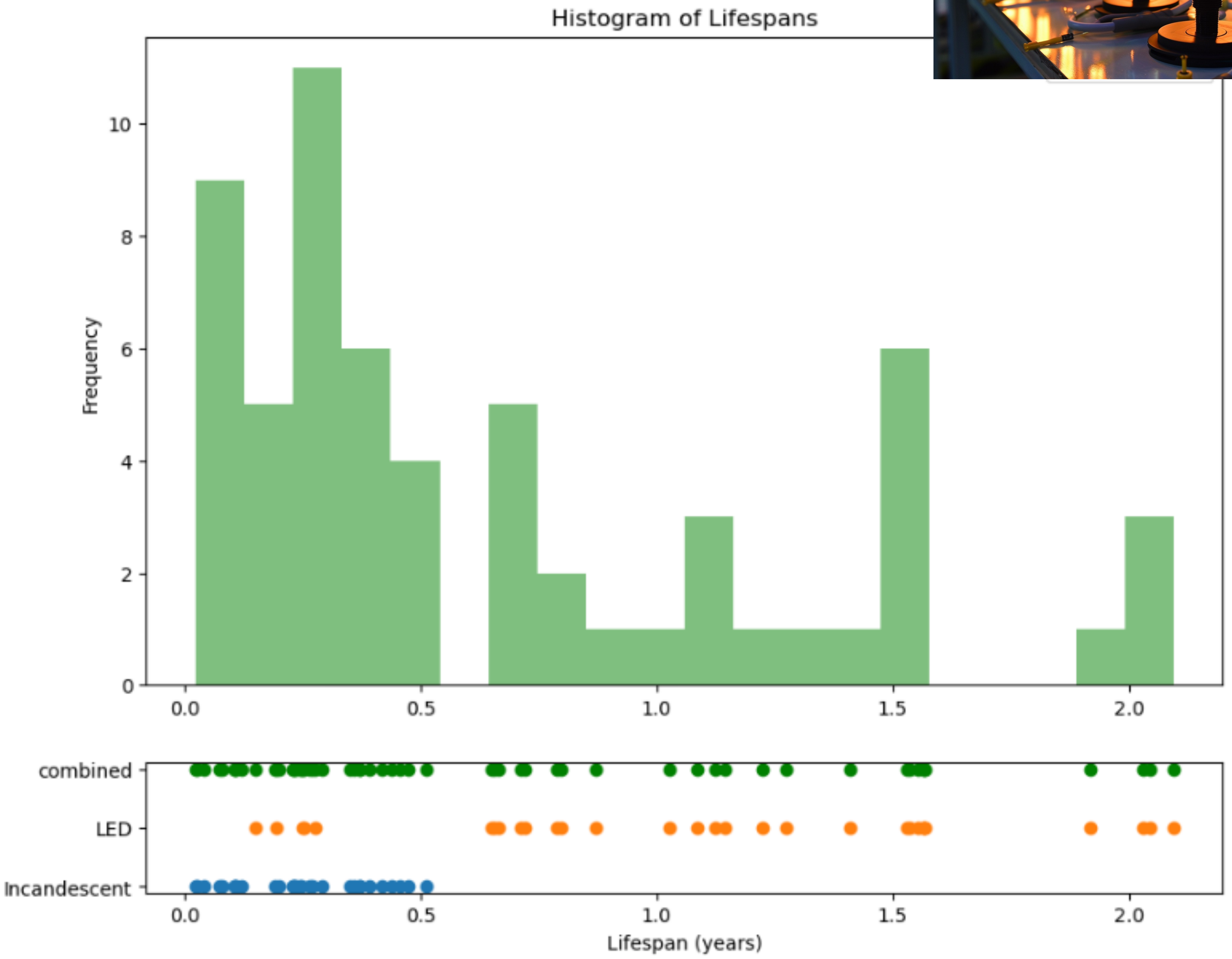


PREVIOUSLY: EXAMPLE: INCANDESCENT VS. LED LIGHTS



Step 1. Combine samples from different groups into a single data set.

PREVIOUSLY: EXAMPLE: INCANDESCENT VS. LED LIGHTS

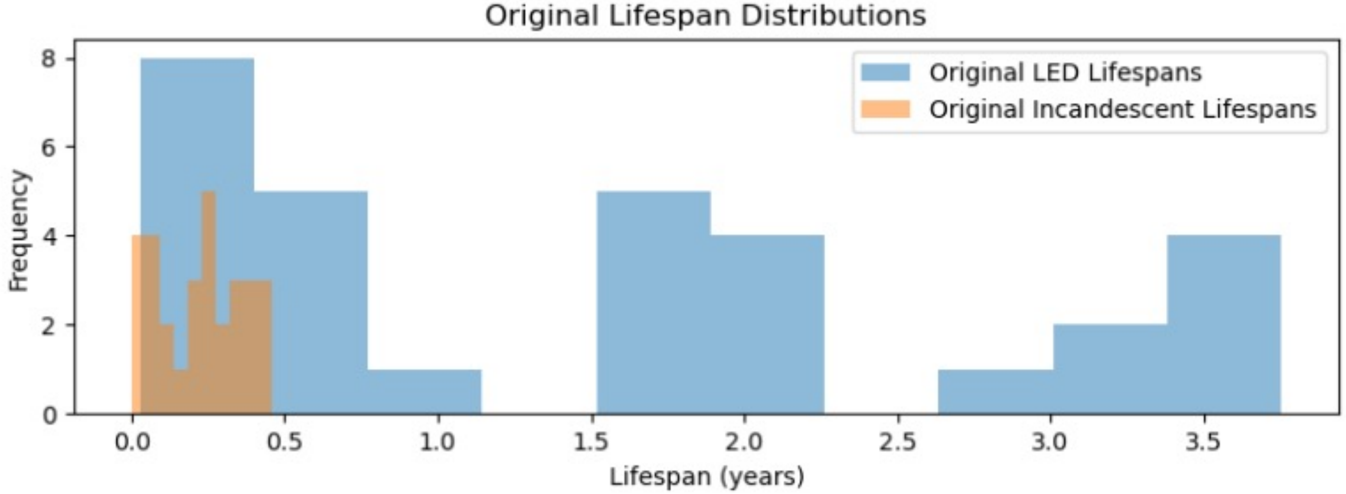


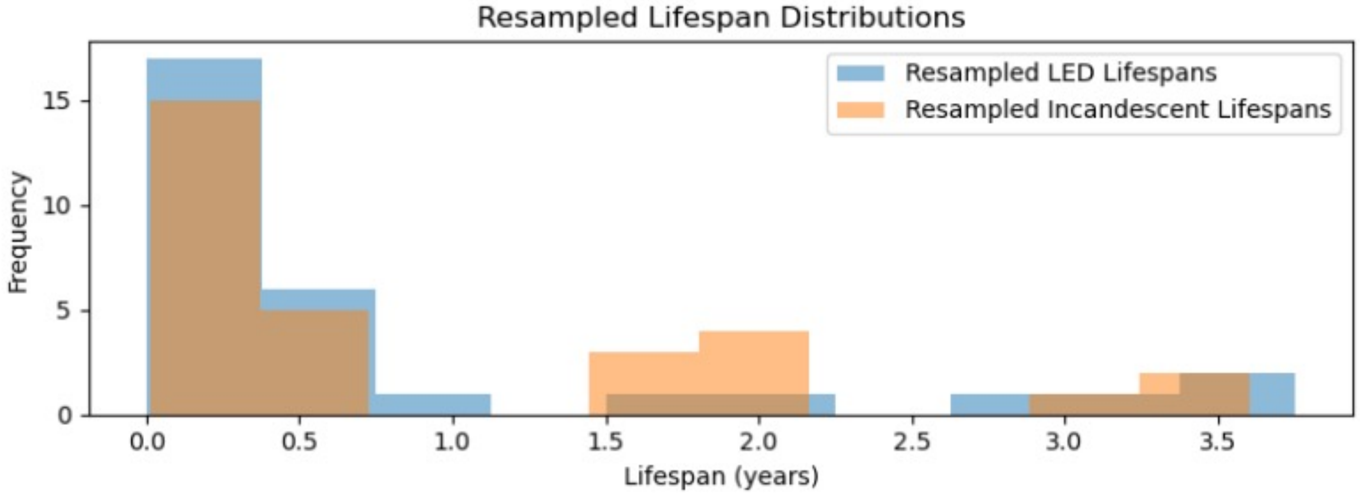
Step 2. Shuffle the combined data set and randomly draw without replacement same size as group A.

PREVIOUSLY: EXAMPLE: INCANDESCENT VS. LED LIGHTS

Step 2. Shuffle the combined data set and randomly draw without replacement same size as group A.

Step 3: Draw without replacement same size as group B

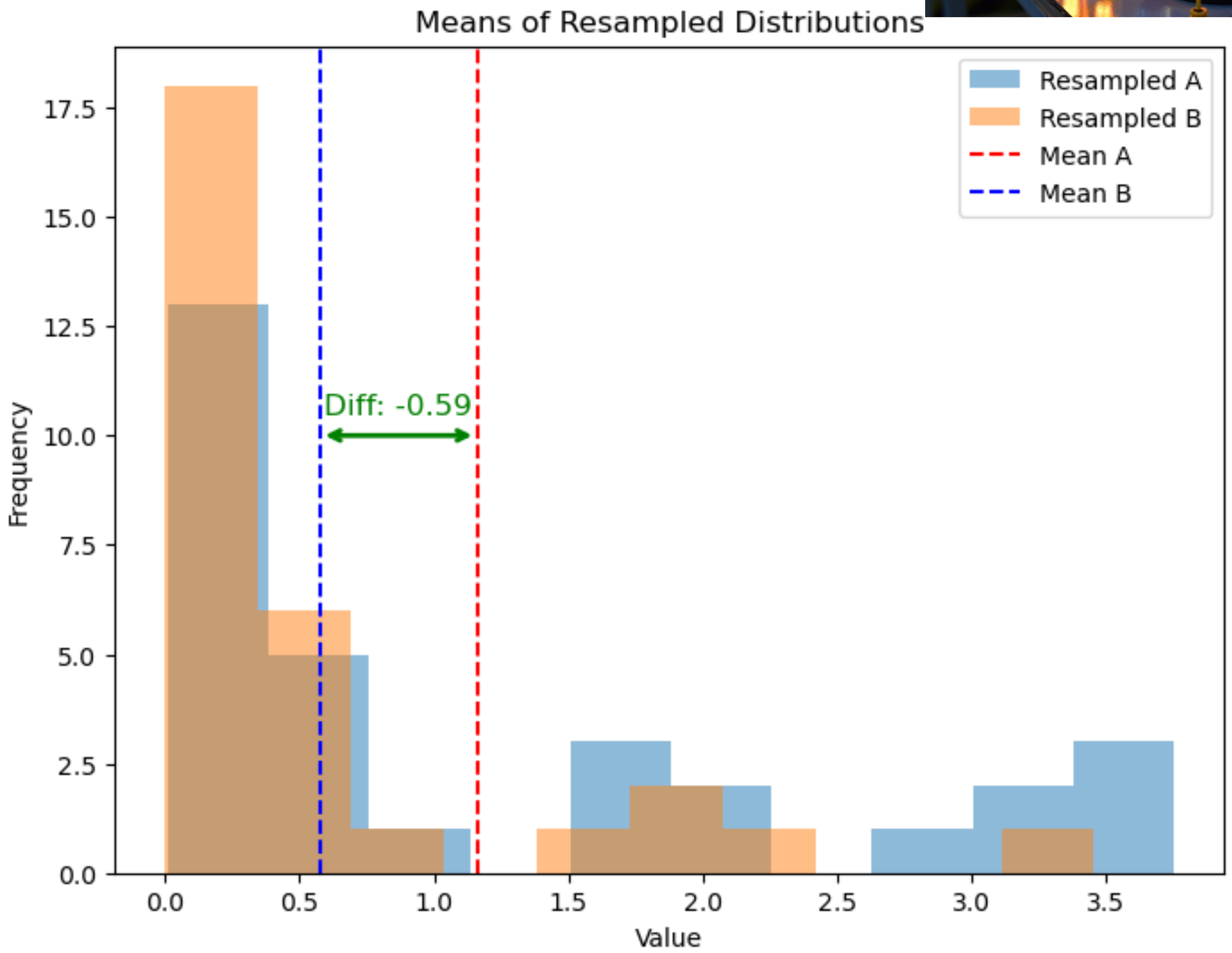




PREVIOUSLY: EXAMPLE: INCANDESCENT VS. LED LIGHTS

Step 4. Measure test statistic.

In this case we are measuring the difference in the means.



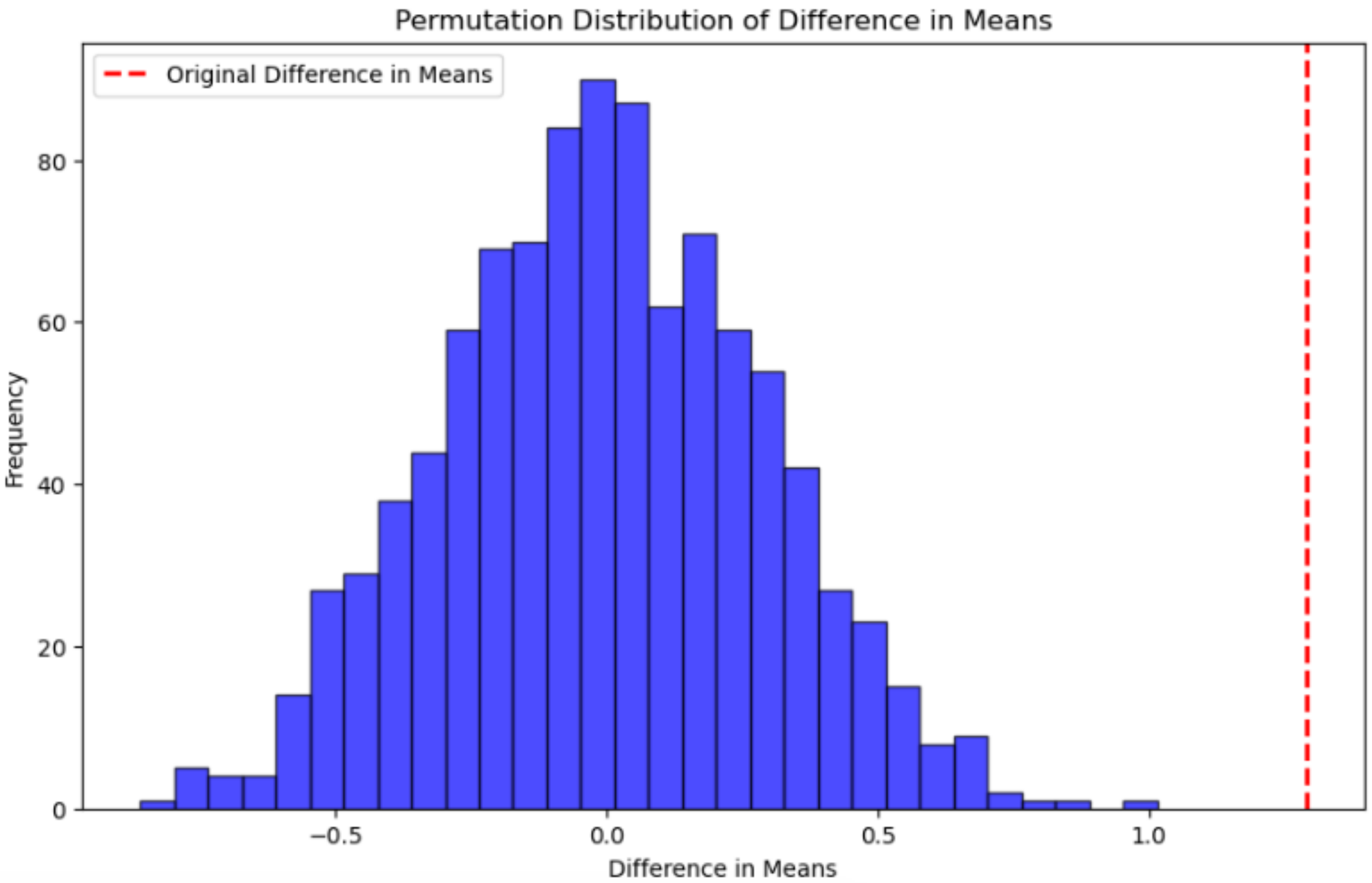
PREVIOUSLY: EXAMPLE: INCANDESCENT VS. LED LIGHTS

Step 5. Repeat until R times to build a permutation distribution.

R=100

A permutation distribution is analogous to a sampling distribution.

It estimates how much our sample means would vary if the null hypothesis is true.



PREVIOUSLY: EXAMPLE: INCANDESCENT VS. LED LIGHTS

We can use the permutation distribution directly to estimate the p-value.

```python
def compute_p_value(permutation_diffs, original_diff):
    # Two-tailed test p-value
    extreme_values = np.abs(permutation_diffs) >= np.abs(original_diff)
    p_value = np.mean(extreme_values)

    return p_value


original_diff = np.mean(led_lifespans_sorted) - np.mean(incandescent_lifespans_sorted)

# Assuming permutation_diffs and original_diff are already defined
p_value = compute_p_value(permutation_diffs, original_diff)

print(f"P-value: {p_value}")
```

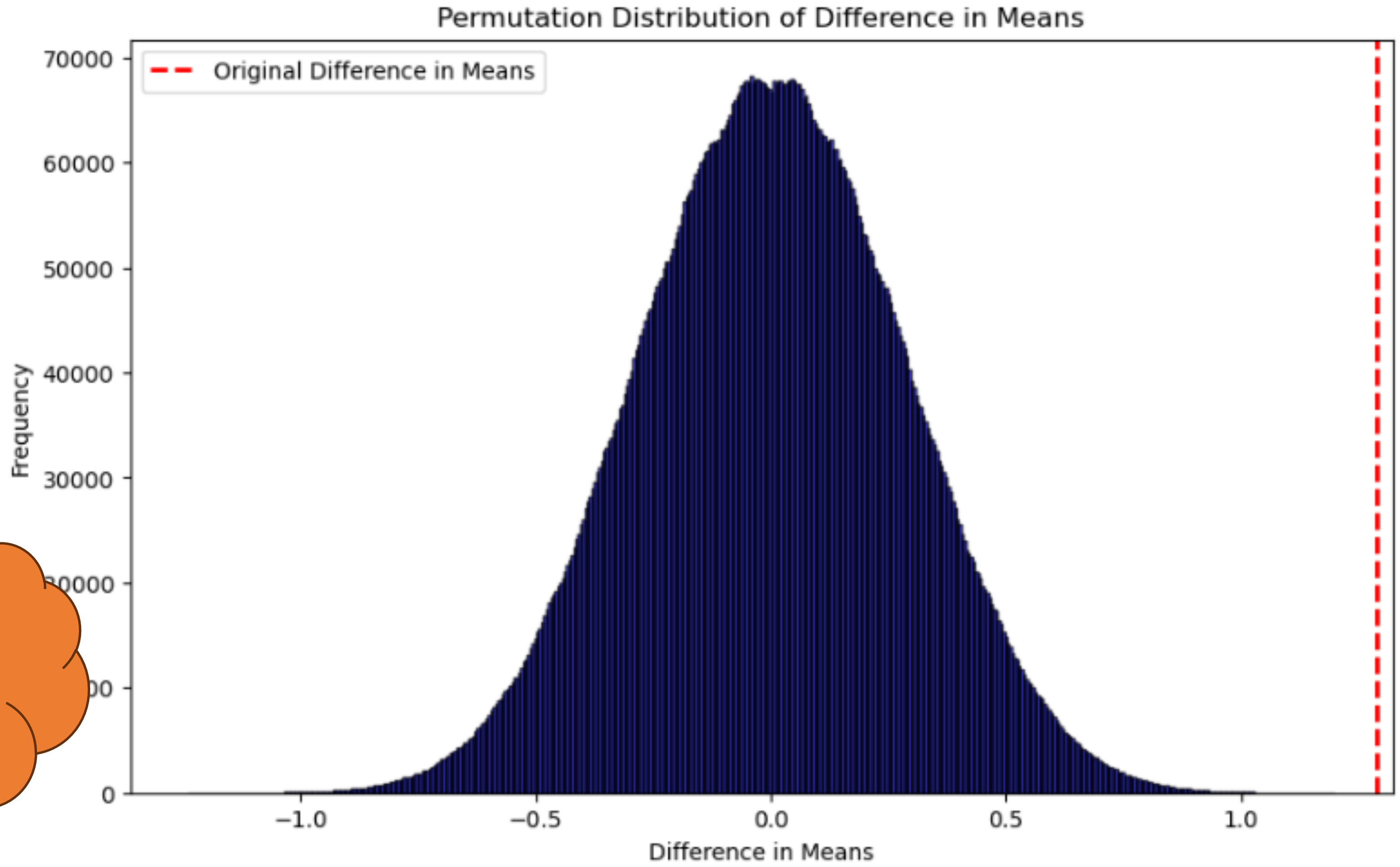
```
P-value: 0.0
```

PREVIOUSLY: EXAMPLE 2: INCANDESCENT VS. LED LIGHTS

Given that the original difference is greater than the differences in all 1000 resamples. We get a p–v of 0.

I reran with 10,000,000 resamples.

```
P-value: 0.0
```



P-value still ZERO!

EXAMPLE: WEB STICKINESS WITH 4 PAGES

Example from the book.

We have four web pages. We swap the page between visitors.

- Each home page has 5 visitors.
- We measure the seconds each visitor spends on the page.

We want to know if there is a difference in stickiness between the pages.

Table 3-3. Stickiness (in seconds) of four web pages

	Page 1	Page 2	Page 3	Page 4
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Average	172	185	176	162
Grand average				173.75

EXAMPLE: WEB STICKINESS WITH 4 PAGES

If we had only two pages A and B we could state a null hypothesis has:

H_0: A and B have the same stickiness.

H_a: A and B DO NOT have the same stickiness.

We could perform A/B hypothesis tests for each pair.

$$\binom{4}{2} = 6 = \text{ number of A/B tests}$$

Table 3-3. Stickiness (in seconds) of four web pages

	Page 1	Page 2	Page 3	Page 4
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Average	172	185	176	162
Grand average				173.75

EXAMPLE: WEB STICKINESS WITH N PAGES

If we have n pages then performing A/B comparisons for each pair would require

$$\binom{n}{2} = \frac{n(n-1)}{2} = O(n^2) = \text{ number of A/B tests}$$

This very rapidly gets unwieldy.

More importantly it has all the problems with multiple hypothesis testing. We would need to adjust α, e.g., using Bonferroni's method or False Discovery Rate method.

Table 3-3. Stickiness (in seconds) of four web pages

	Page 1	Page 2	Page 3	Page 4
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Average	172	185	176	162
Grand average				173.75

EXAMPLE: WEB STICKINESS WITH N PAGES

Instead of pairwise hypothesis tests.

Let’s state a single hypothesis test:

H_0: All pages have the same mean stickiness.

H_a: at least one page has a significantly different mean stickiness.

Table 3-3. Stickiness (in seconds) of four web pages

	Page 1	Page 2	Page 3	Page 4
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Average	172	185	176	162
Grand average				173.75

KEY TERMS FOR ANOVA

Pairwise comparison

A hypothesis test (e.g., of means) between two groups among multiple groups.

Omnibus test

A single hypothesis test of the overall variance among multiple group means.

Decomposition of variance

Separation of components contributing to an individual value (e.g., from the overall average, from a treatment mean, and from a residual error).

F-statistic

A standardized statistic that measures the extent to which differences among group means exceed what might be expected in a chance model.

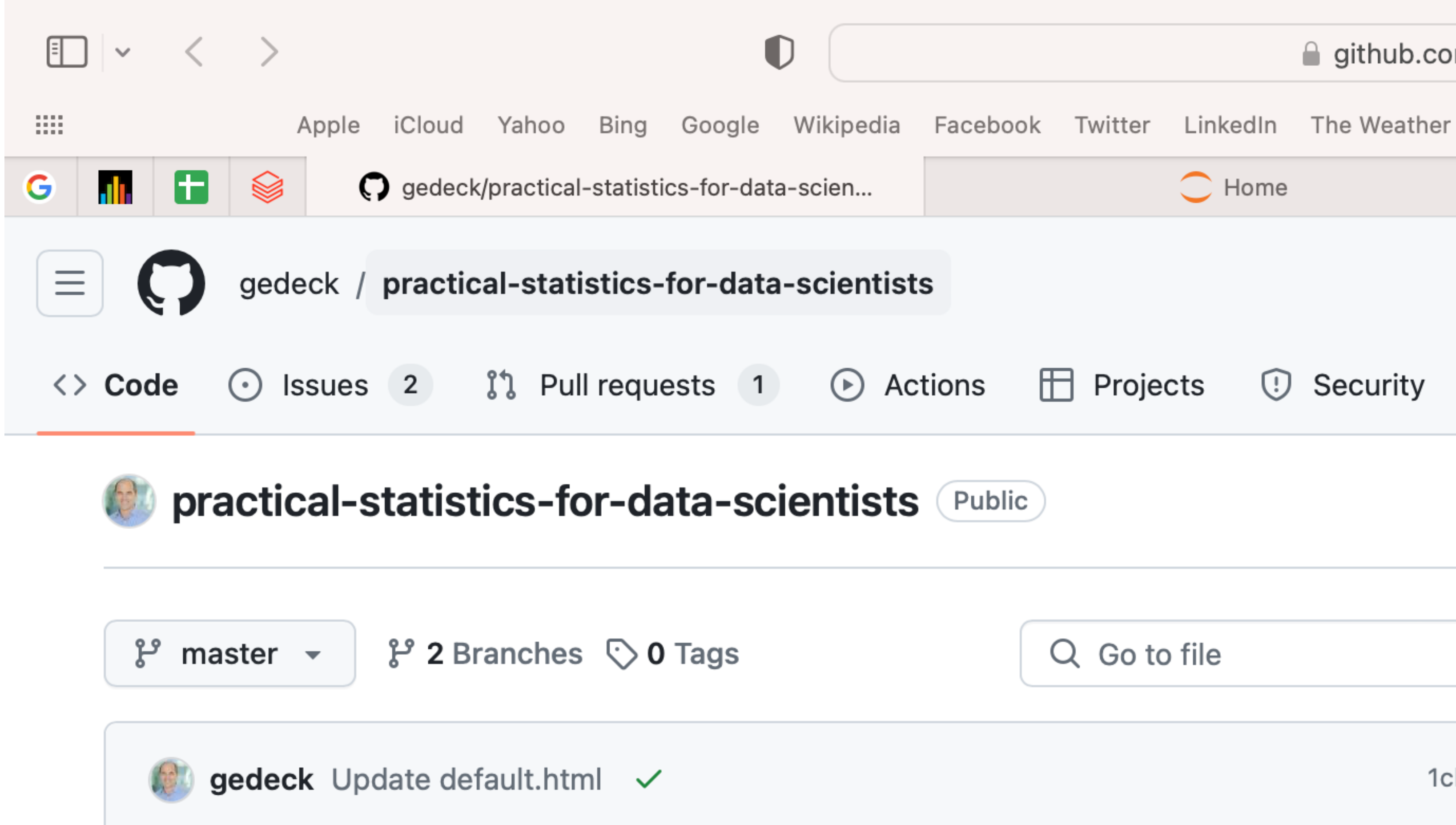
SS

"Sum of squares," referring to deviations from some average value.

EXAMPLE: WEB STICKINESS WITH 4 PAGES

The complete data can be found in the repository for the book.

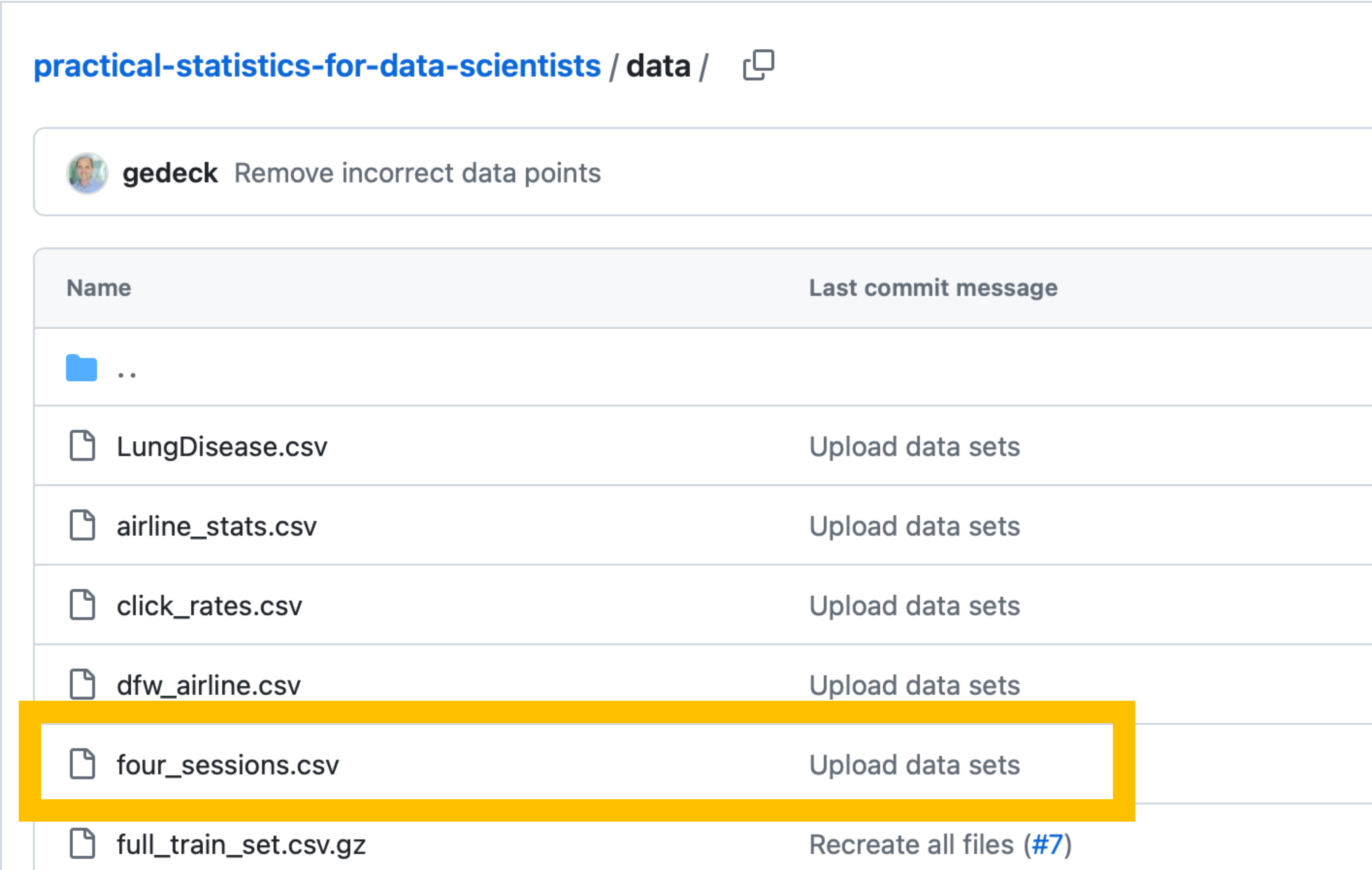
https://github.com/gedeck/practical-statistics-for-data-scientists



EXAMPLE: WEB STICKINESS WITH 4 PAGES

The complete data can be found in the repository for the book.

https://github.com/gedeck/practical-statistics-for-data-scientists/data/four_sessions.csv



EXAMPLE: WEB STICKINESS WITH 4 PAGES

The complete data can be found in the repository for the book.

https://github.com/gedeck/practical-statistics-for-data-scientists/data/four_sessions.csv

Preview Code Blame

```
Page,Time
Page 1,164
Page 2,178
Page 3,175
Page 4,155
Page 1,172
Page 2,191
Page 3,193
Page 4,166
Page 1,177
Page 2,182
Page 3,171
Page 4,164
Page 1,156
Page 2,185
Page 3,163
```

Preview Code Bla

Search this file

1	Page	Time
2	Page 1	164
3	Page 2	178
4	Page 3	175
5	Page 4	155
6	Page 1	172
7	Page 2	191
8	Page 3	193

EXAMPLE: WEB STICKINESS WITH 4 PAGES

```python
import numpy as np
import pandas as pd

# Load the dataset
four_sessions = pd.read_csv('four_sessions.csv')

print(four_sessions.head())

observed_variance = four_sessions.groupby('Page').mean().var().iloc[0]
print('Observed means:', four_sessions.groupby('Page').mean().values.ravel())
print('Variance:', observed_variance)
number_of_rows = len(four_sessions)
print(f'Number of rows (samples): {number_of_rows}')
```

```
     Page  Time
0  Page 1   164
1  Page 2   178
2  Page 3   175
3  Page 4   155
4  Page 1   172
Observed means: [172.8 182.6 175.6 164.6]
Variance: 55.426666666666655
Number of rows (samples): 20
```

Table 3-3. Stickiness (in seconds) of four web pages

	Page 1	Page 2	Page 3	Page 4
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Average	172	185	176	162
Grand average				173.75

EXAMPLE: STICKINESS BOX PLOT

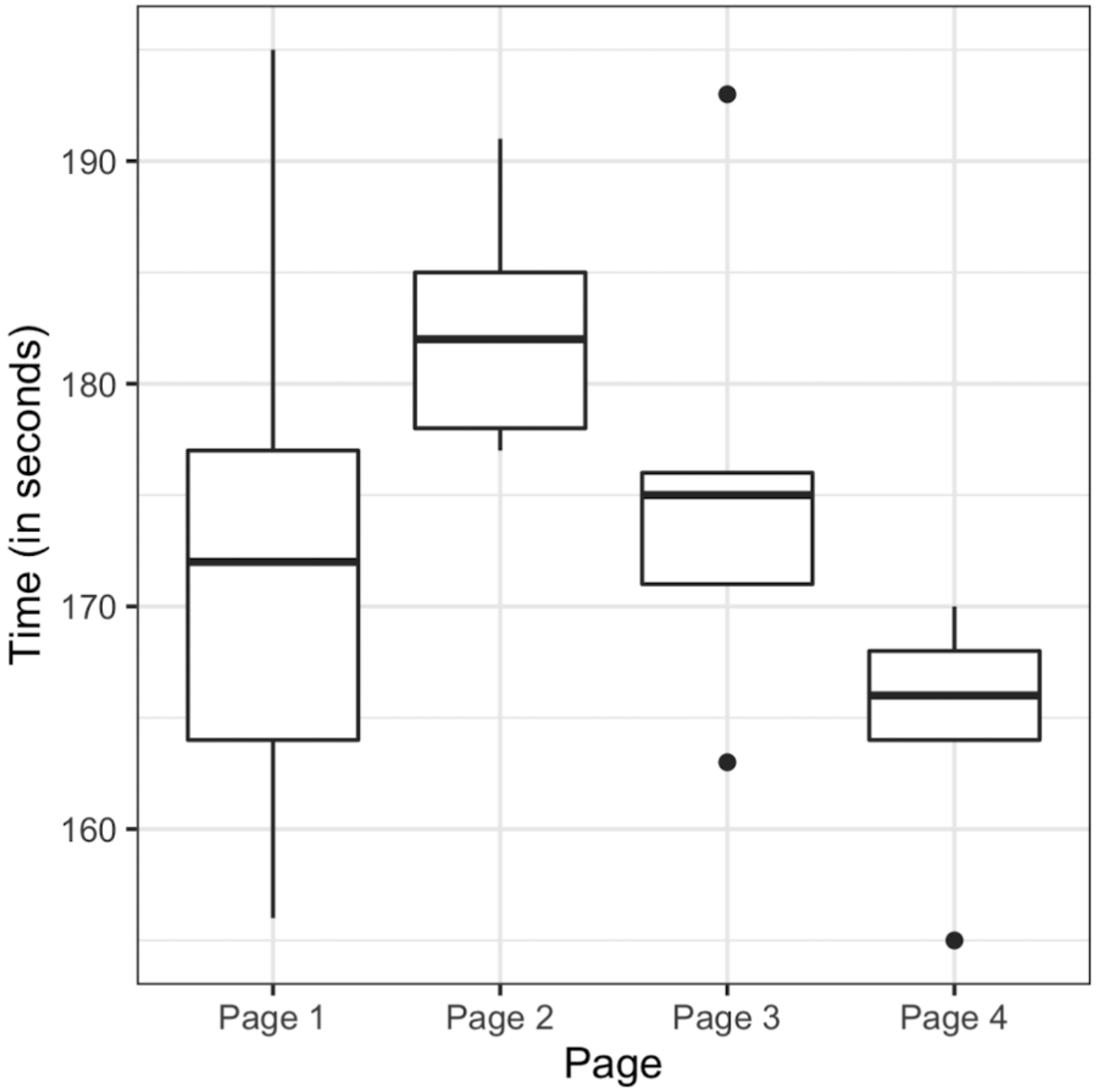


Figure 3-6. Boxplots of the four groups show

EXAMPLE: STICKINESS BOX PLOT

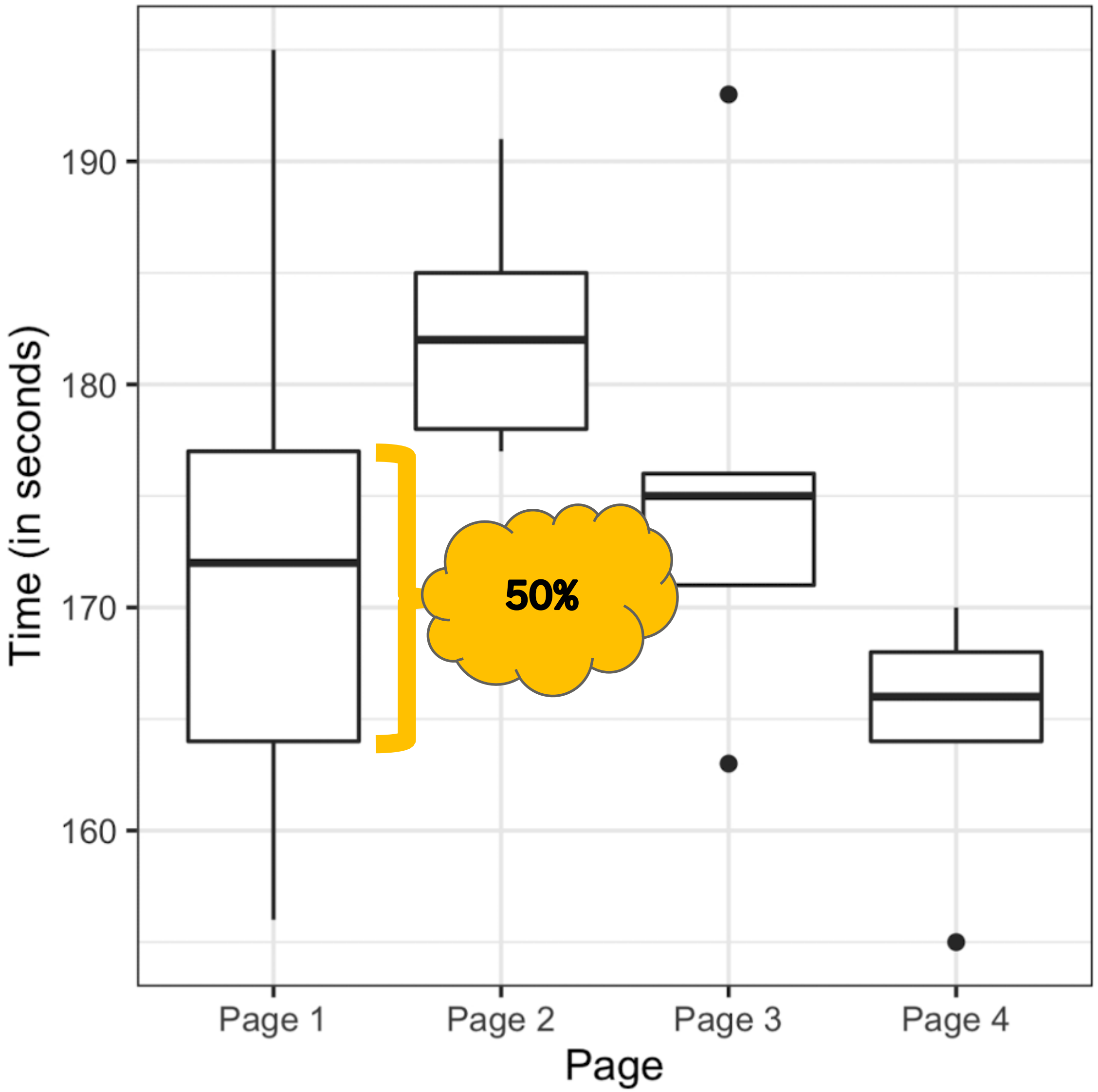


Figure 3-6. Boxplots of the four groups show

EXAMPLE: STICKINESS BOX PLOT

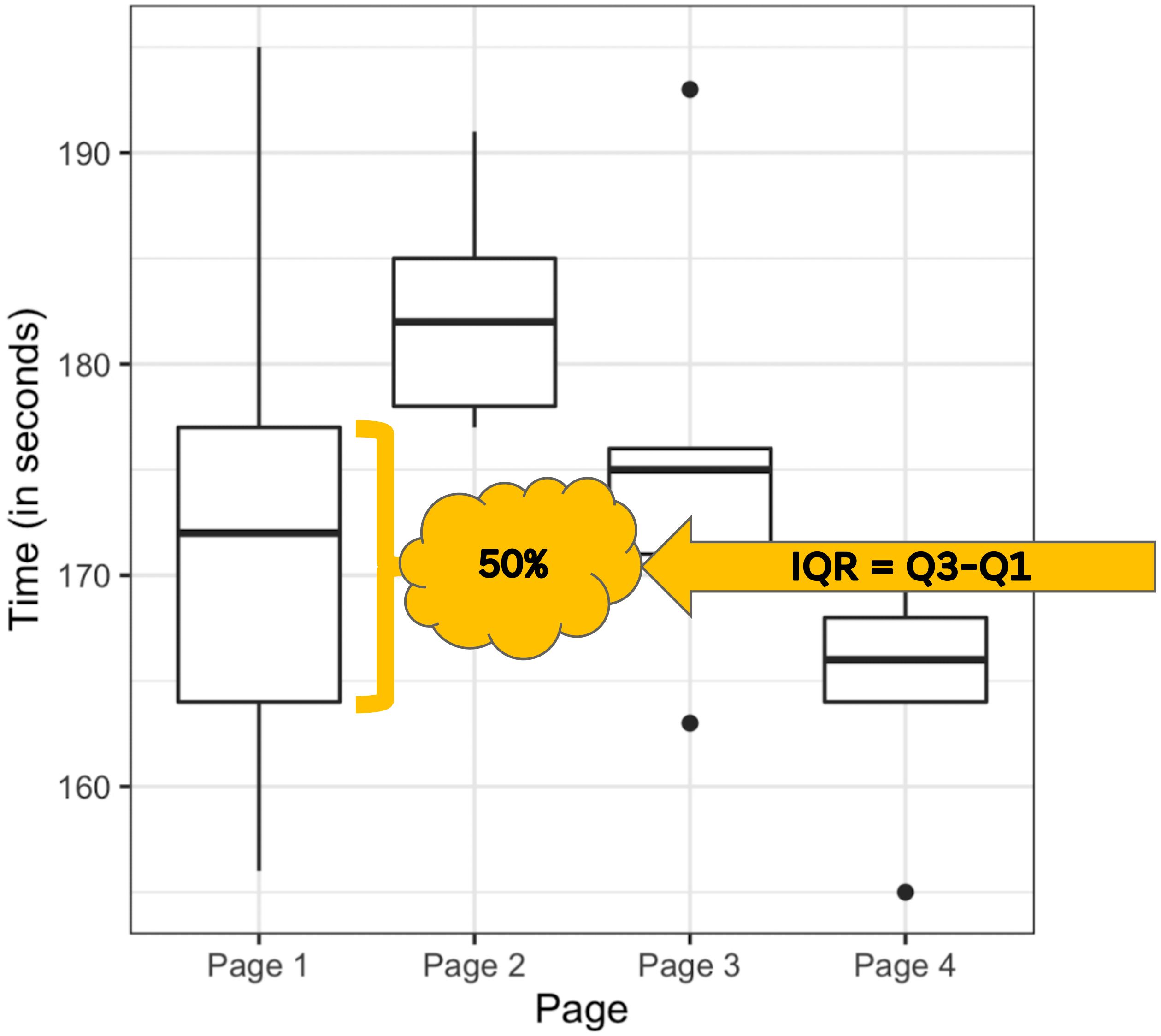


Figure 3-6. Boxplots of the four groups show

EXAMPLE: STICKINESS BOX PLOT

IQR = Interquartile Range

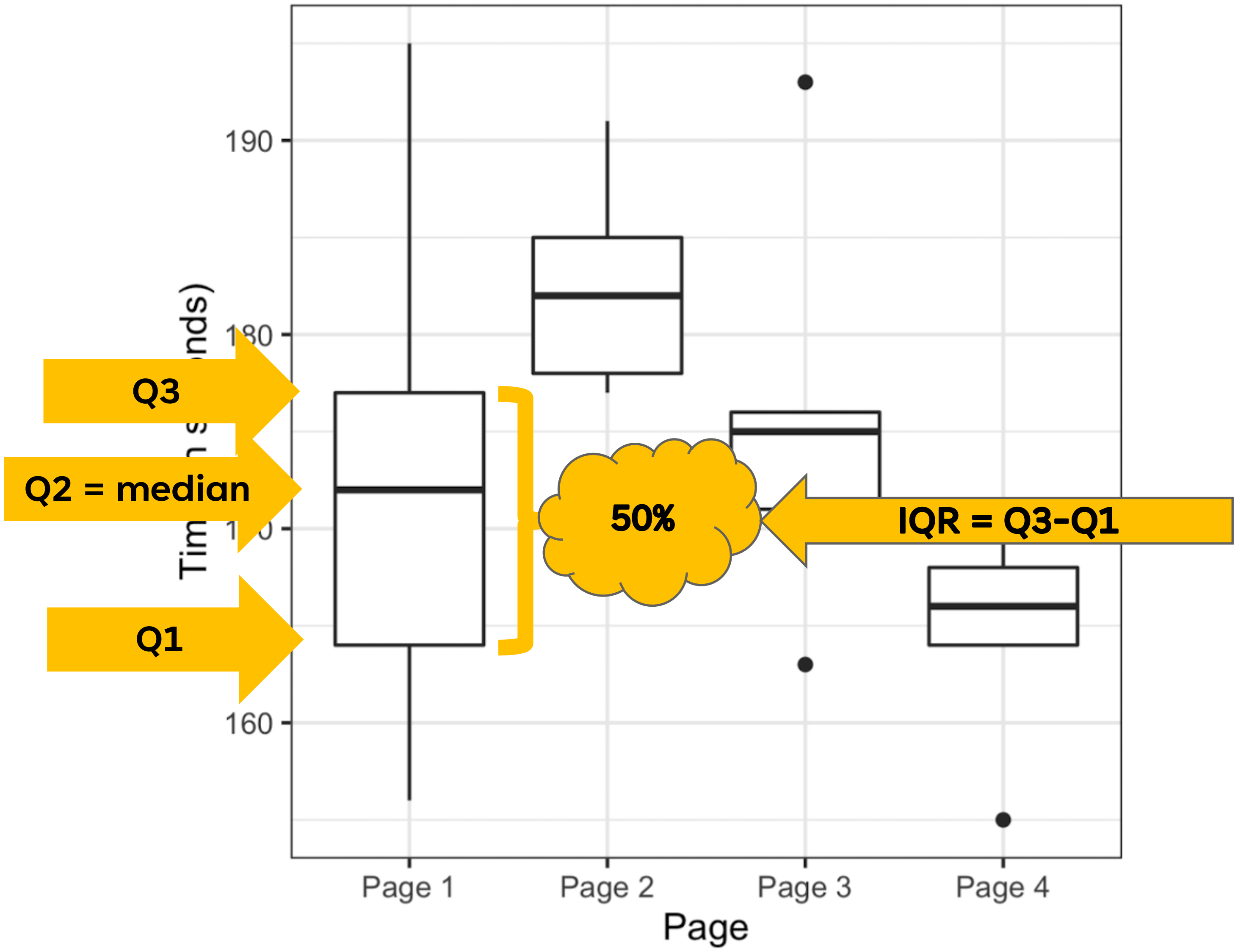


Figure 3-6. Boxplots of the four groups show

EXAMPLE: STICKINESS BOX PLOT

Anything within 1.5 IQR is not considered an outlier.

The lower whisker extends to the lowest value within Q1-1.5 IQR.

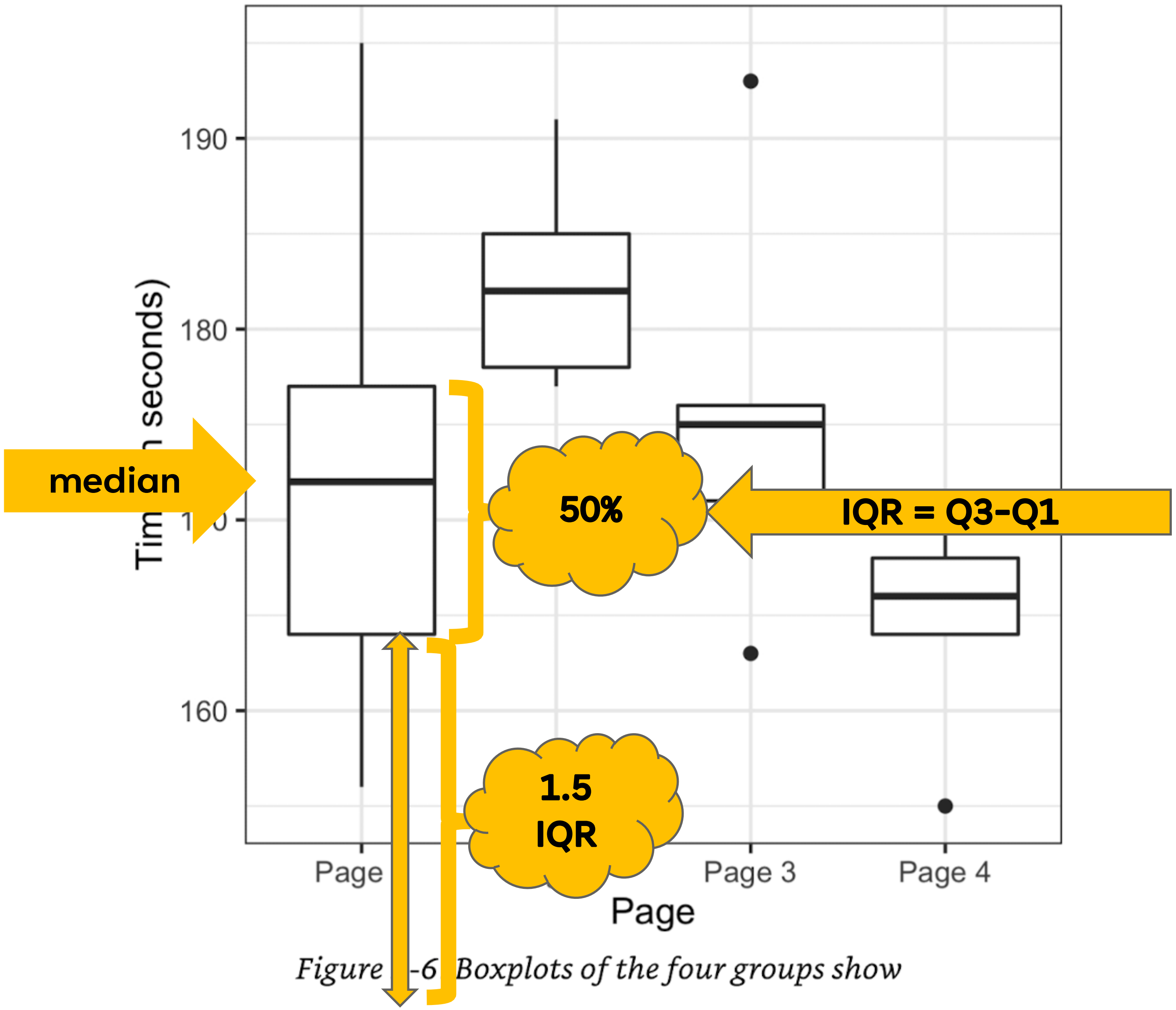


Figure -6 Boxplots of the four groups show

EXAMPLE: STICKINESS BOX PLOT

Anything within 1.5 IQR is not considered an outlier.

The lower whisker extends to the lowest value within Q1-1.5 IQR.

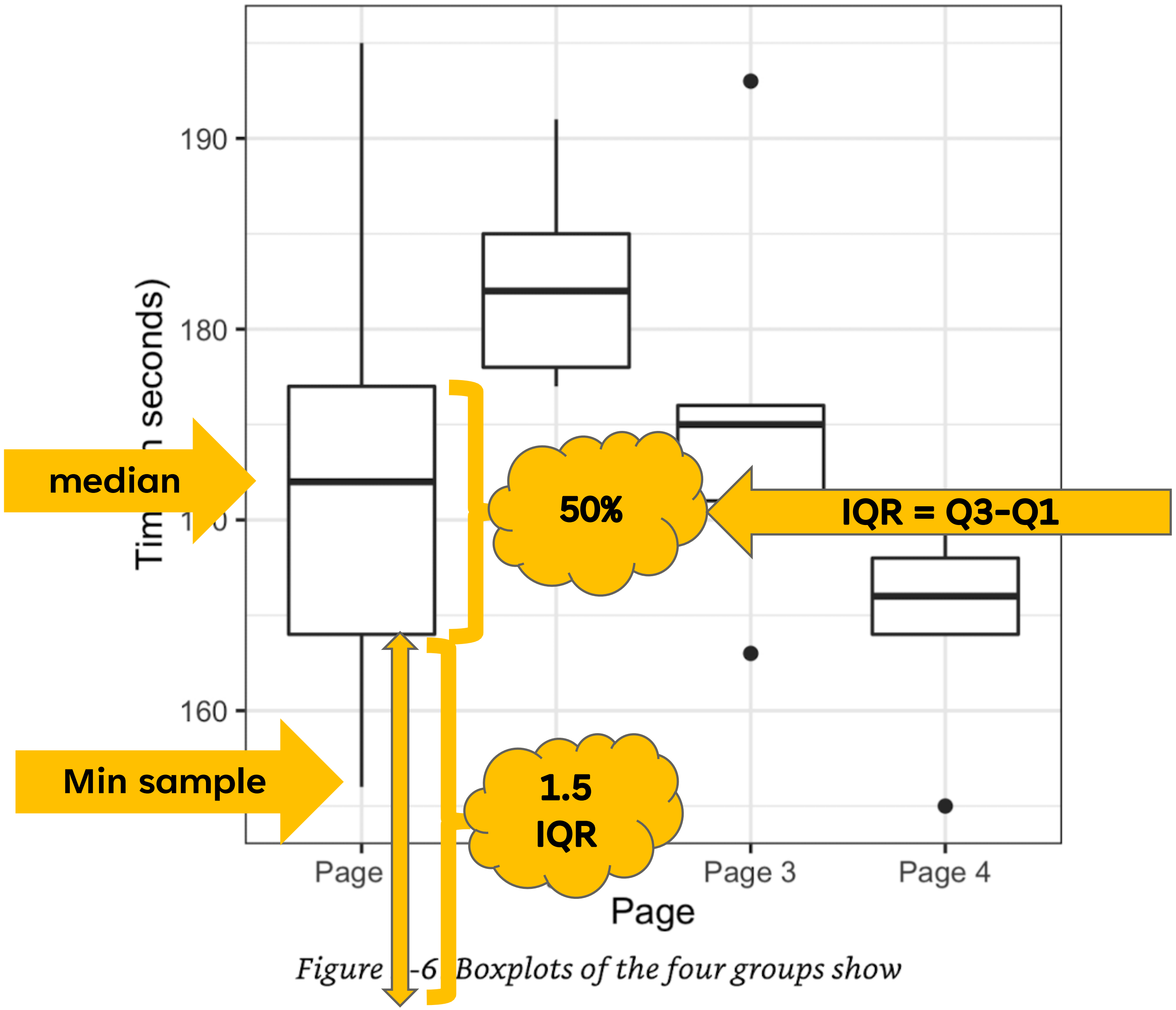


Figure   -6 Boxplots of the four groups show

EXAMPLE: STICKINESS BOX PLOT

Anything within 1.5 IQR is not considered an outlier.

The upper whisker extends to the largest value within Q3 + 1.5 IQR.

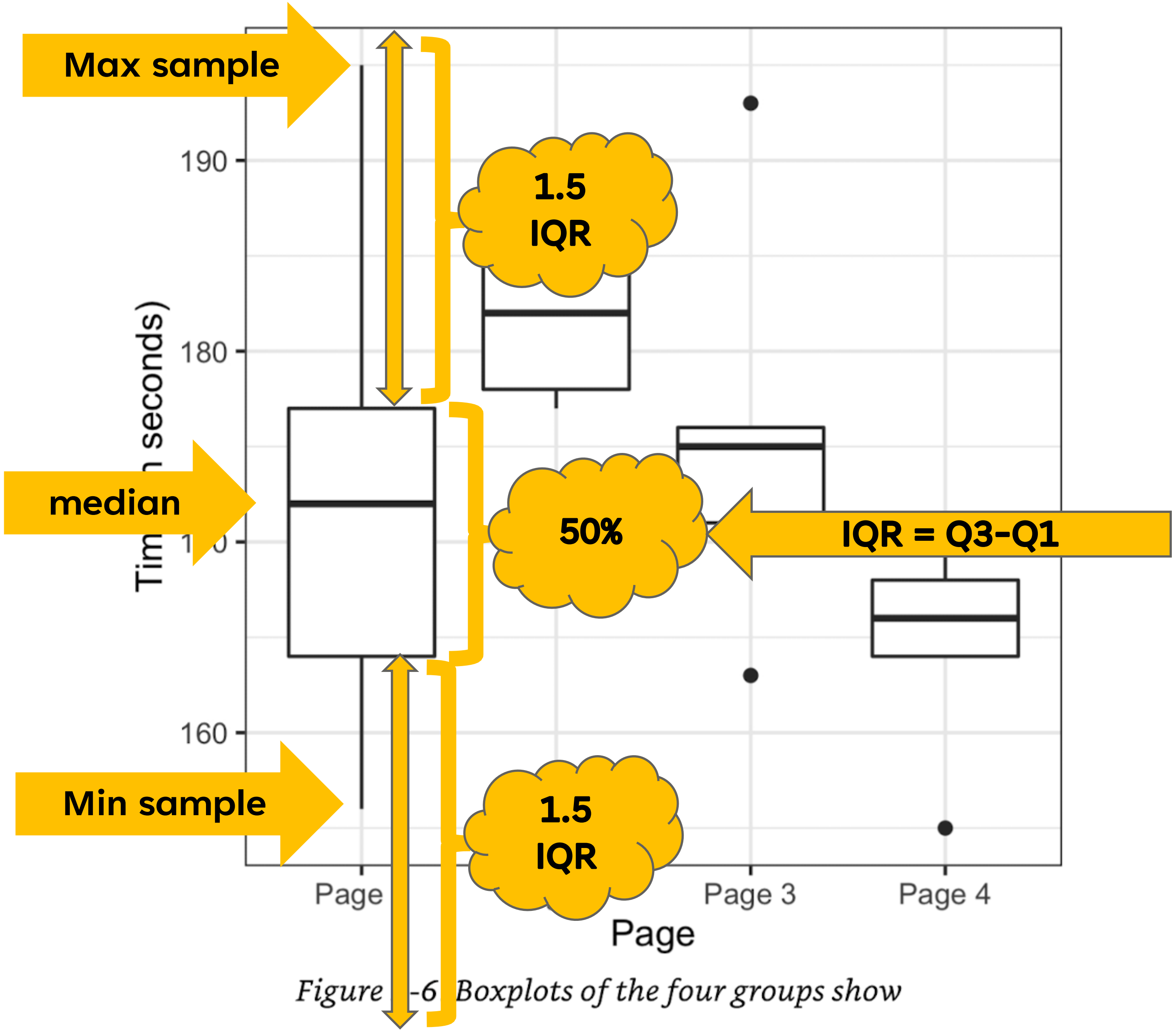


Figure -6 Boxplots of the four groups show

EXAMPLE: STICKINESS BOX PLOT

Why no whiskers?

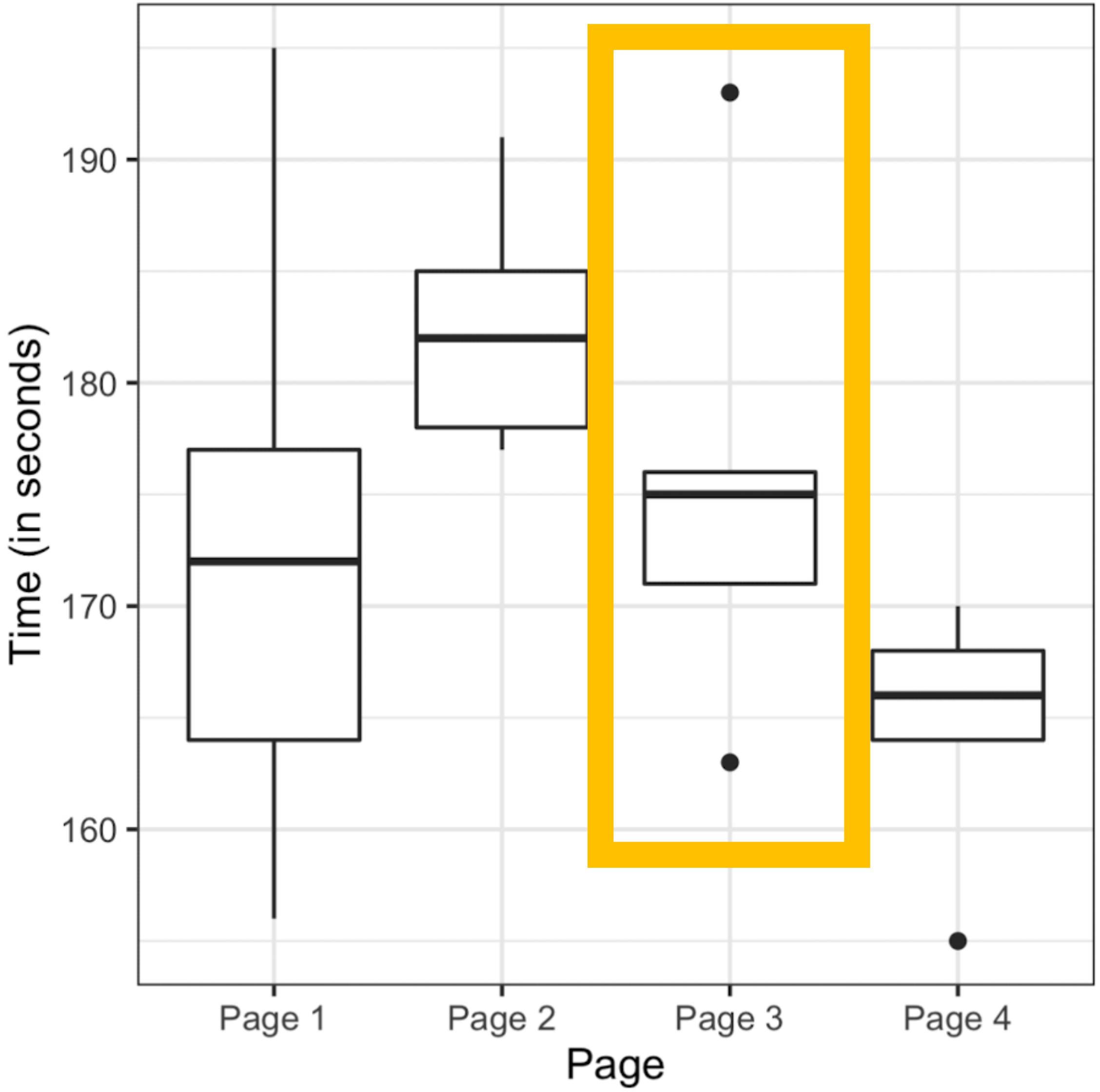


Figure 3-6. Boxplots of the four groups show

EXAMPLE: STICKINESS BOX PLOT

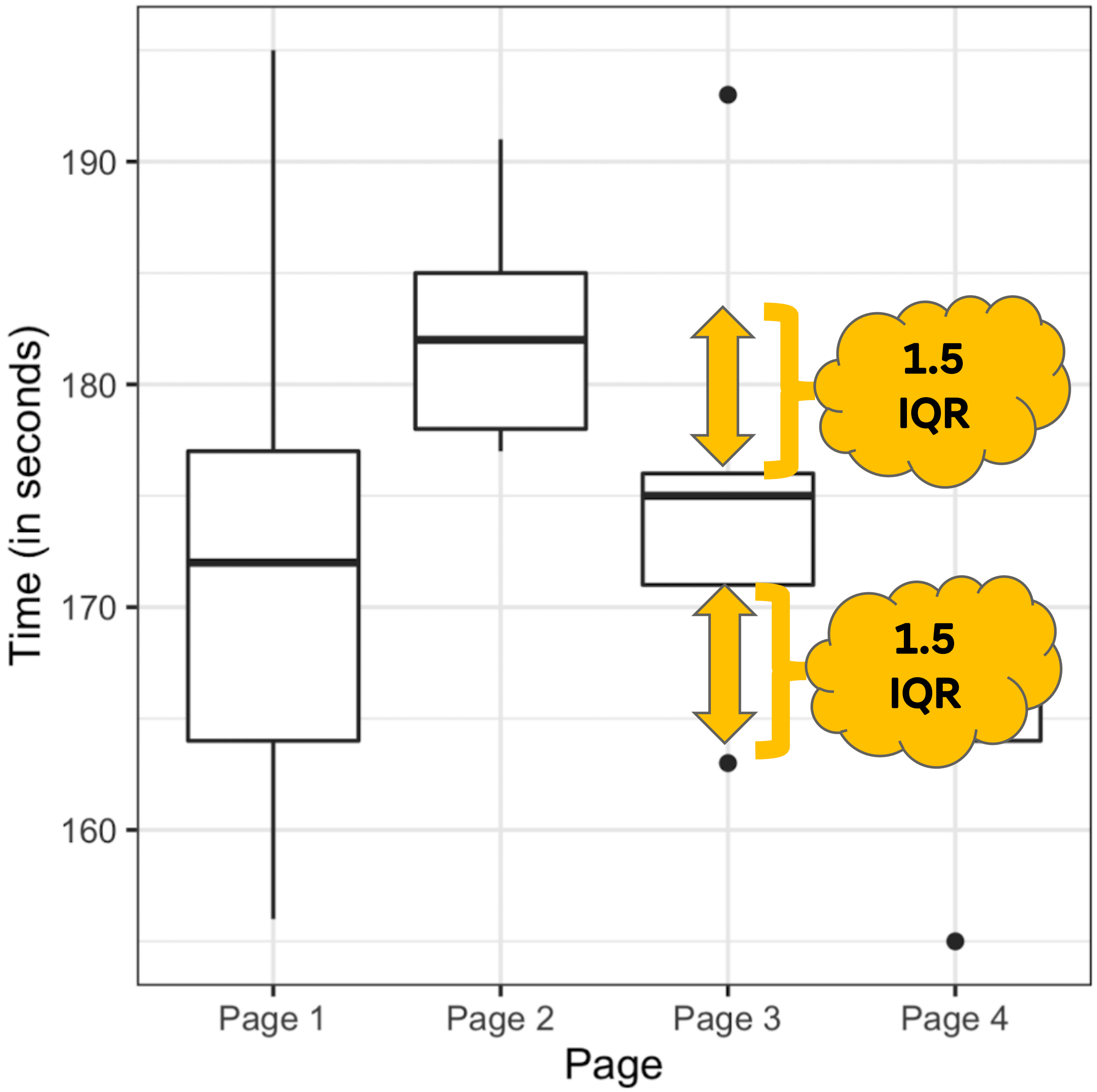


Figure 3-6. Boxplots of the four groups show

Why no whiskers?

A: Because the nearest data points outside the IQR (the box) are not within 1.5 IQR.

EXAMPLE: STICKINESS BOX PLOT

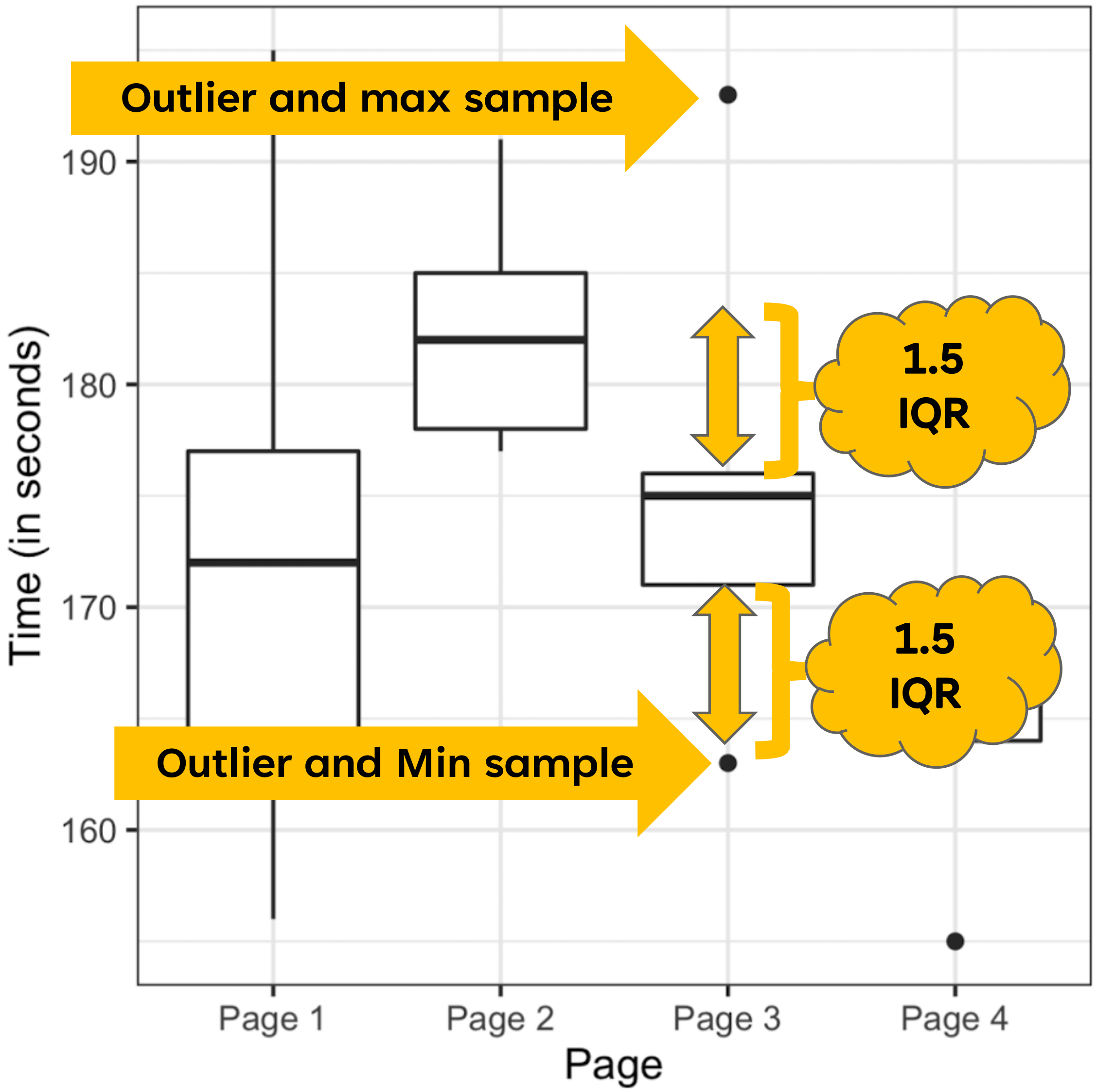


Figure 3-6. Boxplots of the four groups show

EXAMPLE: STICKINESS BOX PLOT

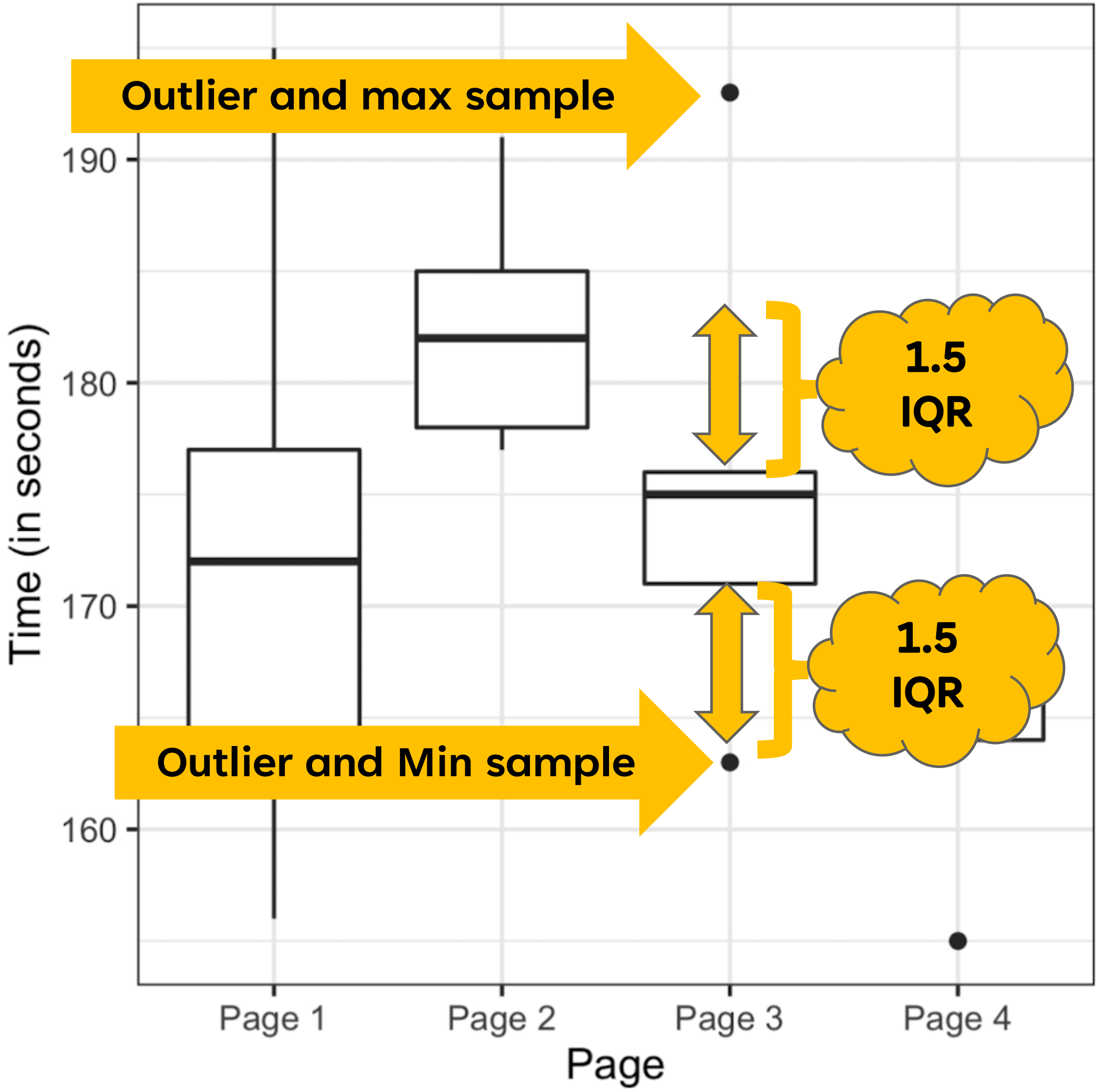


Figure 3-6. Boxplots of the four groups show

EXAMPLE: STICKINESS BOX PLOT

Mixed.

Max is within 1.5 IQR of Q3.

Min is outside 1.5 IQR of Q1.

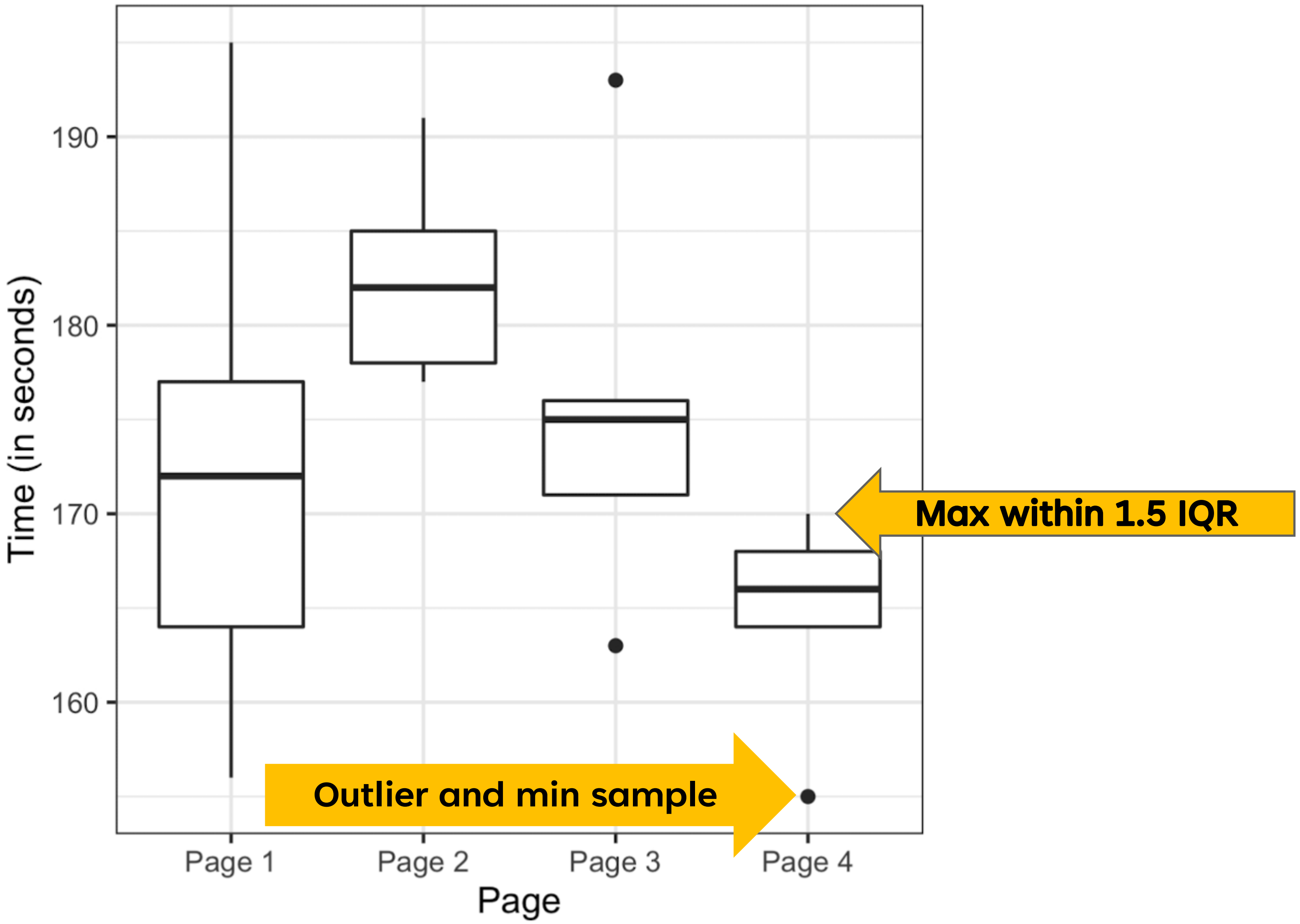


Figure 3-6. Boxplots of the four groups show

EXAMPLE: STICKINESS BOX PLOT

What can we way about the four pages?

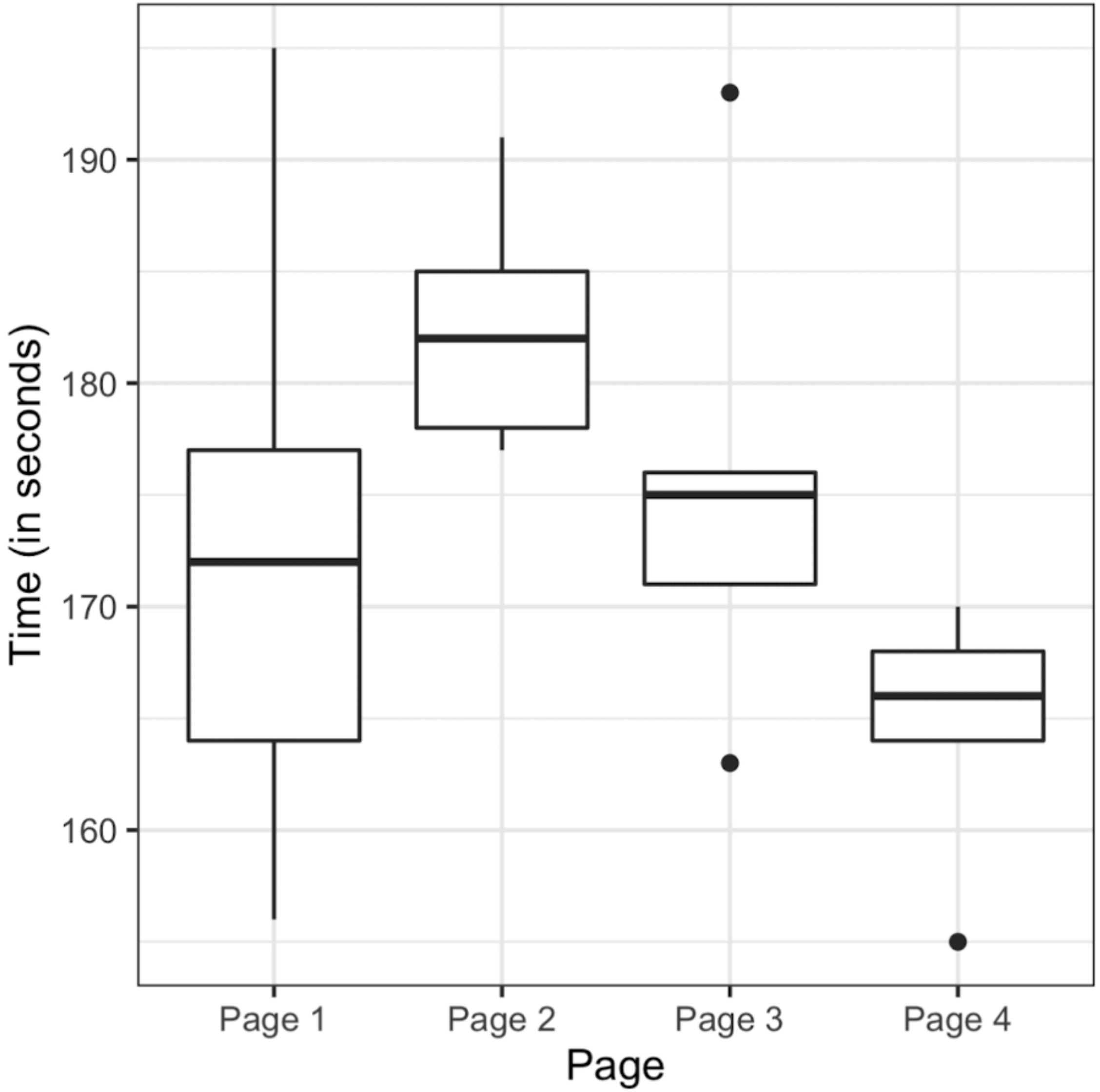


Figure 3-6. Boxplots of the four groups show

EXAMPLE: STICKINESS BOX PLOT

We could say

1. there is a fair amount of overlap in the interquartile ranges.
2. Some but few outliers.
3. Little but not much skew
4. Likely too few samples.

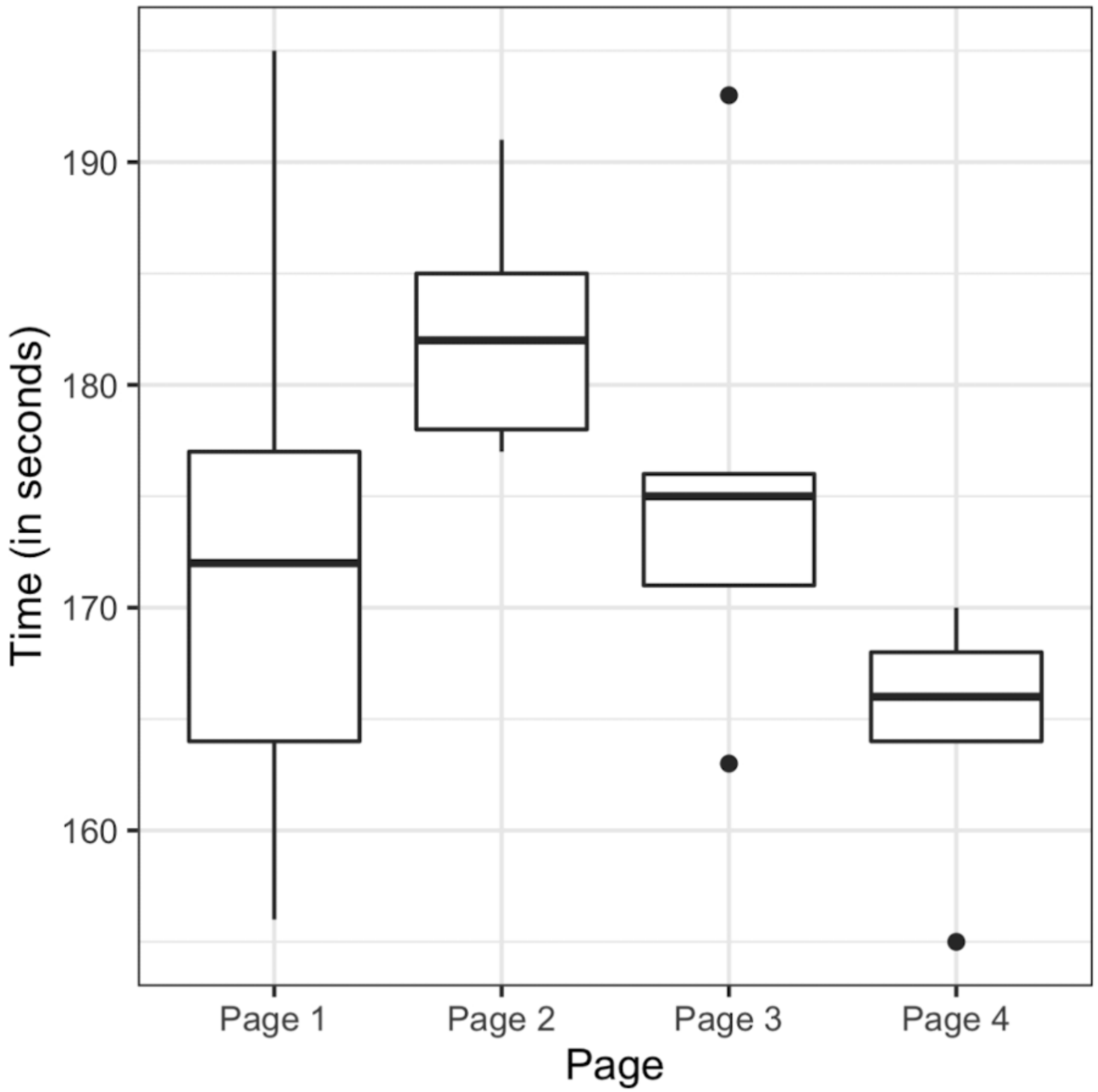


Figure 3-6. Boxplots of the four groups show

EXAMPLE: STICKINESS BOX PLOT

We could say

1. there is a fair amount of overlap in the interquartile ranges.
2. Some but few outliers.
3. Little but not much skew
4. Likely too few samples.
5. Maybe page 2 has significantly greater mean?

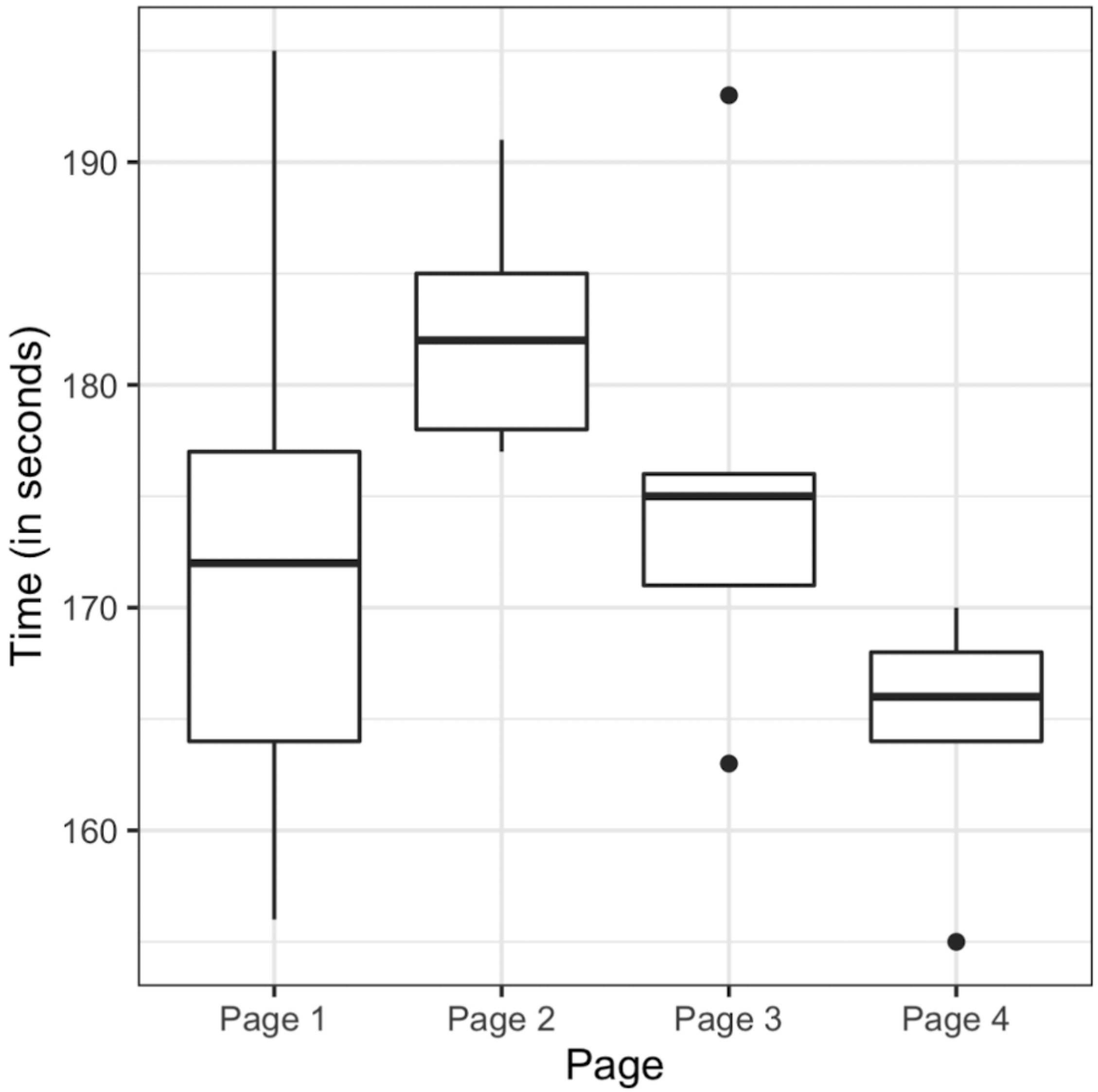


Figure 3-6. Boxplots of the four groups show

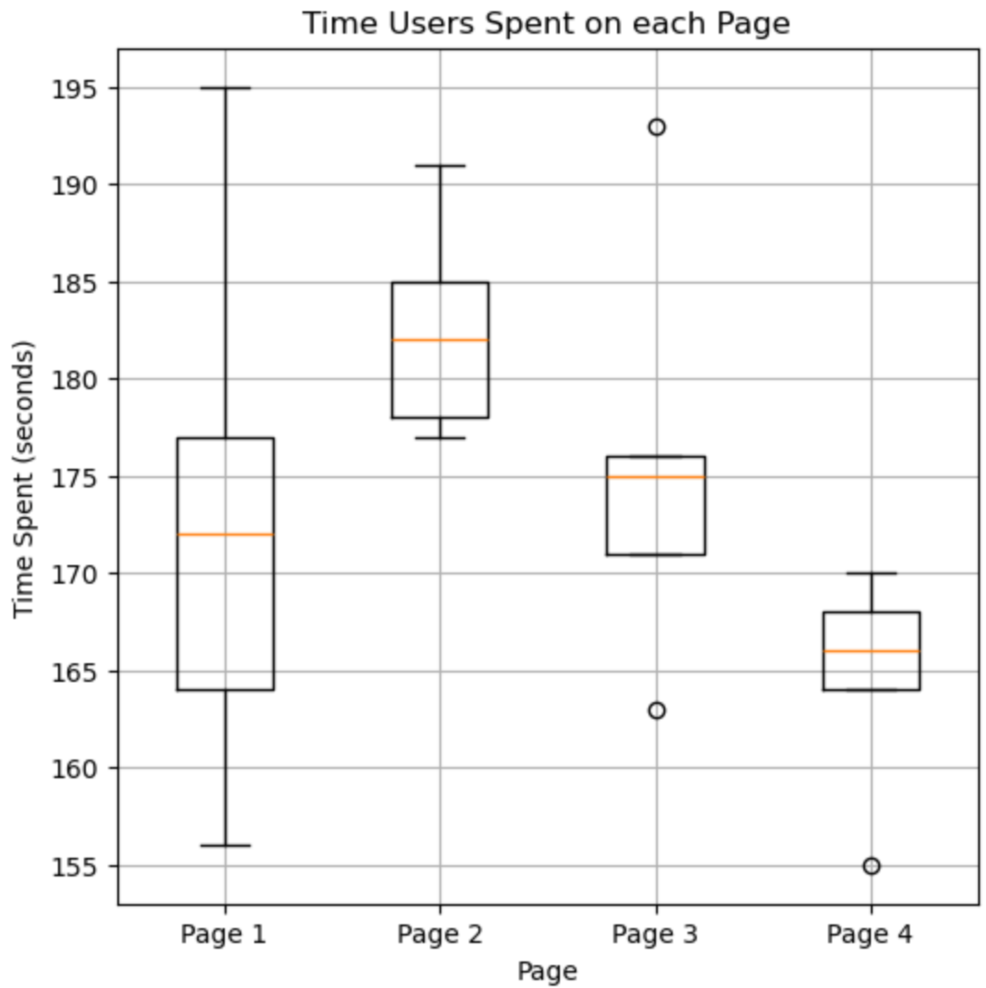
EXAMPLE: STICKINESS BOX PLOT

Generate my own
box and whisker plot

```
import pandas as pd
import matplotlib.pyplot as plt

# Assuming four_sessions is already loaded into your environment
# You may need to load your data as shown previously if starting a new session

# Create a boxplot
plt.figure(figsize=(10, 6))
plt.boxplot([four_sessions[four_sessions['Page'] == page]['Time'] for page in four_sessions['Page'].unique()],
            labels=four_sessions['Page'].unique())
plt.title('Box and Whisker Plot of Time on Different Pages')
plt.xlabel('Page')
plt.ylabel('Time Spent (seconds)')
plt.grid(True)
plt.show()
```



EXAMPLE: STICKINESS BOX PLOT

Generate my own
box and whisker plot

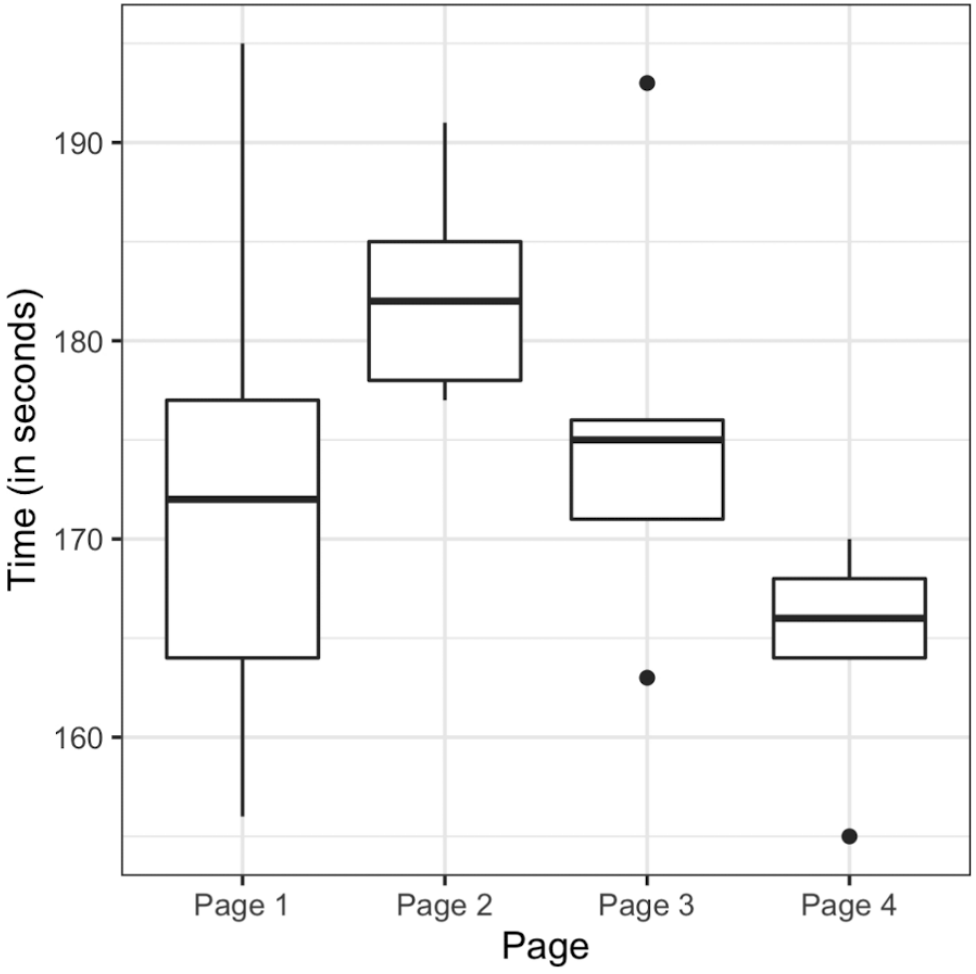
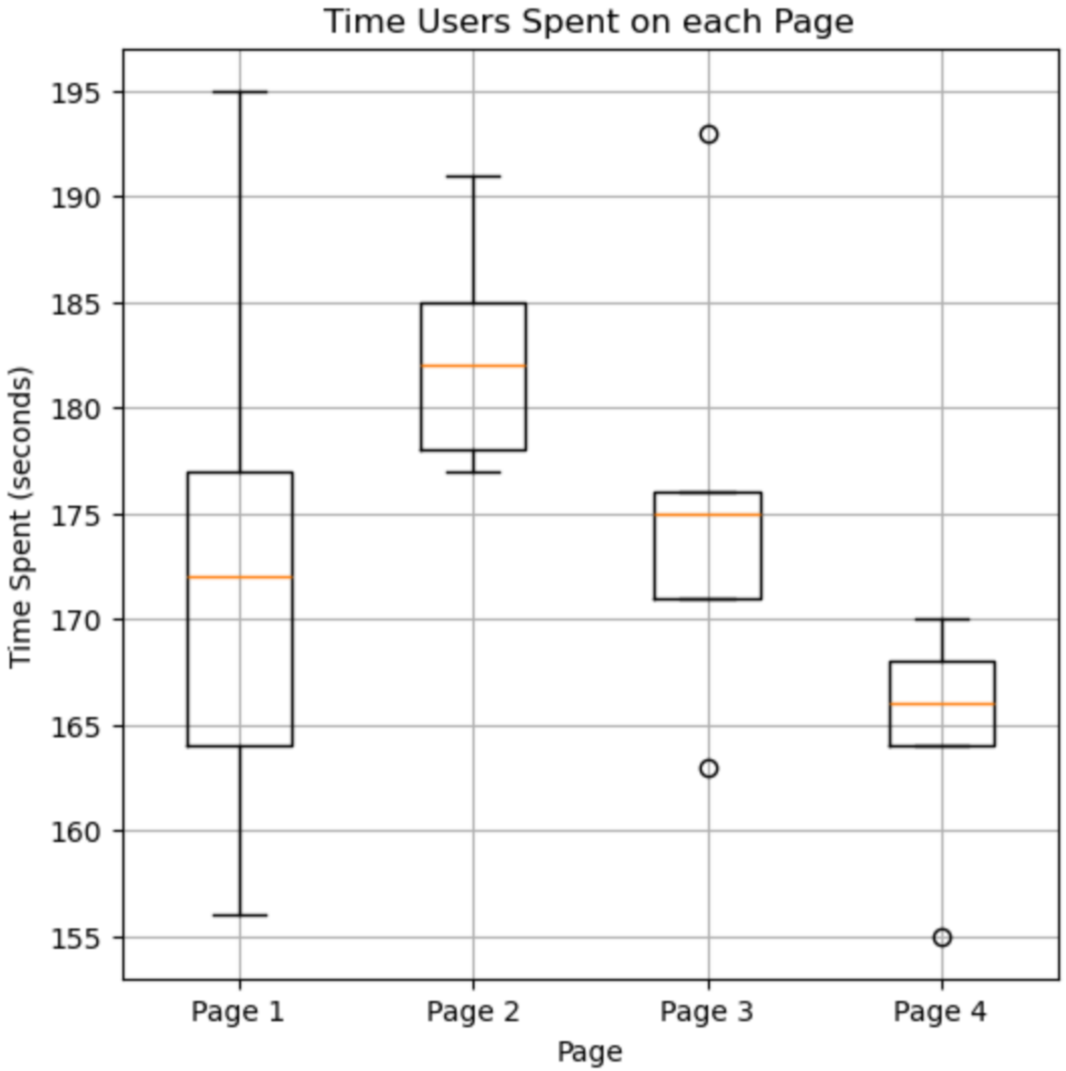


Figure 3-6. Boxplots of the four groups show



TRADITIONAL ANOVA

ANOVA = ANalysis Of Variance

- Given multiple groups (presumed to be independent).
- Null hypothesis (H_0): "The group means are NOT different."
- Alternative hypothesis (H_a): "At least one group has different mean."

- **Gaussian**: Assumes distributions for each group are Gaussian.
- **Homoscedasticity**: Assumes the groups have equal variances.
- **Independence**: Assumes the samples in each group are independent of each other.

- If sample size is large enough Gaussian may be satisfied due to CLT.
- If sample variances are similar, then homoscedasticity is satisfied.

WHAT IF ASSUMPTIONS ARE VIOLATED?

As with A/B tests, we could use permutation testing.

Permutation ANOVA

- Groups may NOT be Gaussian.
- Groups may have unuqual variances.
- **Independence**: for resampling we still assume the samples are independent from each other.

PERMUTATION ANOVA

1. Combine all the data together in a single box.
2. Shuffle and draw out four resamples of five values each.
3. Record the mean of each of the four groups.
4. Record the variance among the four group means.
5. Repeat steps 2–4 many (say, 1,000) times.

1. Combine all data into a single box.
   - Compute the grand mean, mean of all samples.

$$\bar{x}_{\text{grand}} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

2. Shuffle and resample 1 group without replacement for each of the original groups.
   - For each of the original groups we have one resampled group of the same size.
3. Record the mean of each resampled group.

$$\bar{x}_{\text{Page 1}}, \bar{x}_{\text{Page 2}}, \bar{x}_{\text{Page 3}}, \bar{x}_{\text{Page 4}}$$

4. Record the variance among the means .

$$\text{Variance} = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x}_i - \bar{x}_{\text{grand}})^2$$

5. Repeat steps 2-4 many (e.g., R=1000) times.

PERMUTATION ANOVA

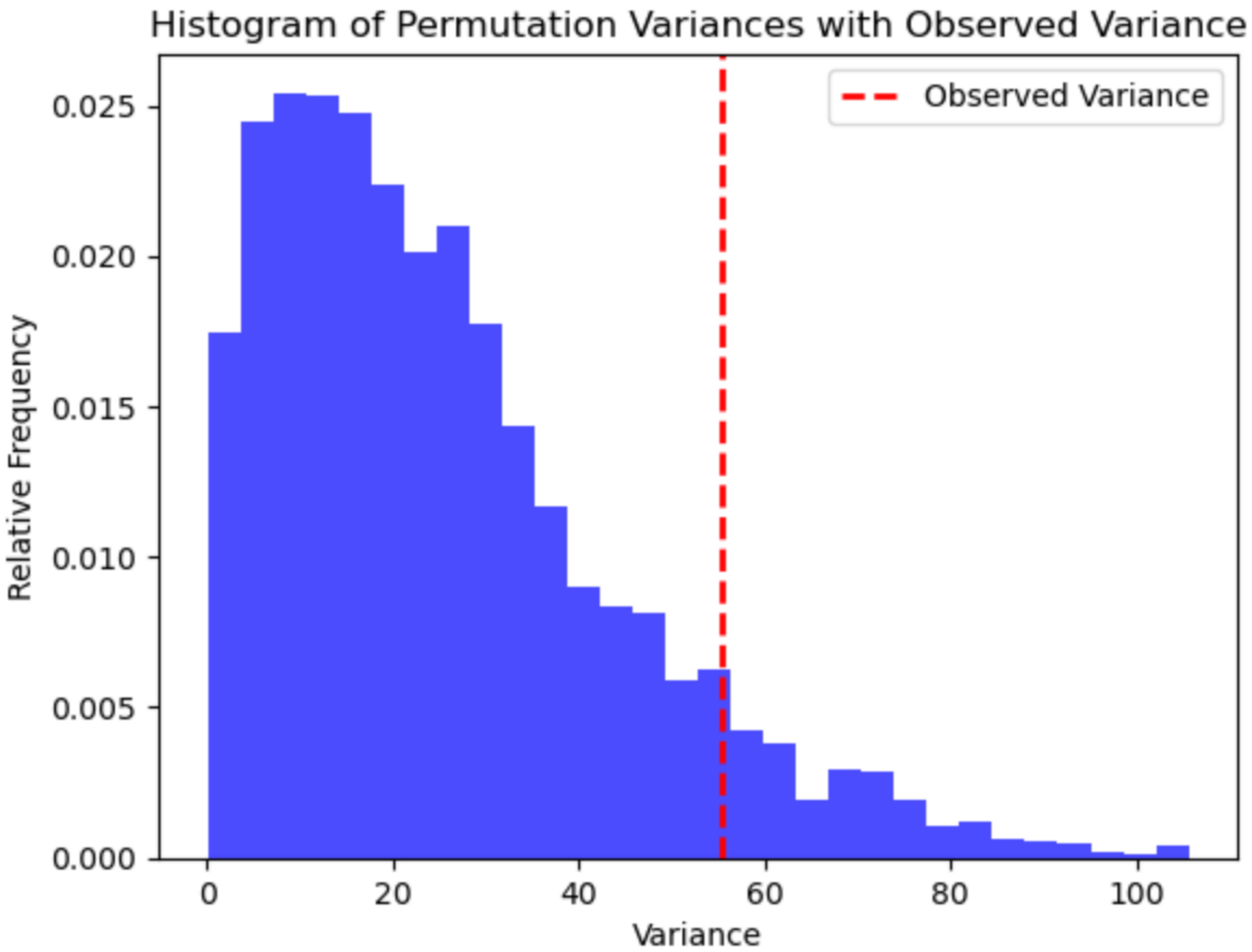
- After this process we have many computed sample variances.
- These variances estimate the distribution of the combined groups.
- As with permutation tests applied to A/B tests, we can evaluate the significance from the permutation distribution.
  - How often did the variance among the resampled group means exceed the variance of the group means in the original data.
  - This is your p-value.
- If the p-value is small this means that resampled variances are unusually large compared to the original variance of the group

```
def perm_test(df):
    df = df.copy()
    df['Time'] = np.random.permutation(df['Time'].values)
    return df.groupby('Page').mean().var().iloc[0]

perm_variance = [perm_test(four_sessions) for _ in range(3000)]
print('P-value', np.mean([var > observed_variance for var in perm_variance]))
```

```
P-value 0.077
```

PERMUTATION ANOVA



PERMUTATION ANOVA

```python
def perm_test(df):
    df = df.copy()
    df['Time'] = np.random.permutation(df['Time'].values)
    return df.groupby('Page').mean().var().iloc[0]

perm_variance = [perm_test(four_sessions) for _ in range(3000)]
print('P-value', np.mean([var > observed_variance for var in perm_variance]))
```

```
P-value 0.077
```

Preview | Code | Blame

```
Page,Time
Page 1,164
Page 2,178
Page 3,175
Page 4,155
Page 1,172
Page 2,191
Page 3,193
Page 4,166
Page 1,177
Page 2,182
Page 3,171
Page 4,164
Page 1,156
Page 2,185
Page 3,163
Page 4,170
Page 1,195
```

1. Copy the DataFrame so that we don’t alter the original data.
2. Permutation randomly shuffles the page stickiness times.
3. `groupby` groups by the “Page” column.
4. `.mean().var()` computes the means and then the variance of the group means.
5. We call `perm_test` 3000 times.
6. Count the number of variances greater than the `observed_variance` (variance of means in original data) and divide by the number of resamples.

PERMUTATION ANOVA

```
def perm_test(df):
    df = df.copy()
    df['Time'] = np.random.permutation(df['Time'].values)
    return df.groupby('Page').mean().var().iloc[0]

perm_variance = [perm_test(four_sessions) for _ in range(3000)]
print('P-value', np.mean([var > observed_variance for var in perm_variance]))
```

```
P-value 0.077
```

The p-value estimates the P[between group variance > observed_variance] directly from the empirical distribution.

```
print ([var > observed_variance for var in perm_variance][:10])
print(f"np.mean([True, False, False]) = {np.mean([True, False, False])}")
```

```
[False, False, False, False, False, False, False, False, False, False]
np.mean([True, False, False]) = 0.3333333333333333
```

Assuming an α=0.05, a p-value of 0.077 is larger than α, so we lack sufficient evidence to reject the null hypothesis. The means might be the same.

THANK YOU

David Harrison

Harrison@cs.olemiss.edu