Final Project

Tutorial section 0212, group 212-3

Zenghao Wang, Oliver Bassel, Eunice Lee, Hyejeong Lee

Introduction

Riipen is a service where students, professors, and companies can collaborate and complete tasks for each other through “requests”. The main dataset we will be working with:

```
## Observations: 2,526
## Variables: 8
## $ Id                <int> 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 16, 18…
## $ Recipient.Id      <int> 1142, 910, 1108, 1108, 910, 1052, 1187, 1161, …
## $ Actor.Id          <int> 18026, 17140, 11839, 16196, 11947, 17730, 9427…
## $ Requestable.Model <fct> project, project, project, project, project, p…
## $ Day.of.Created.At <fct> "12 April, 2018", "12 April, 2018", "12 April,…
## $ Day.of.Updated.At <fct> "24 April, 2018", "18 April, 2018", "2 May, 20…
## $ Day.of.Expired.At <fct> "", "", "", "", "", "", "6 December, 2018", ""…
## $ State             <fct> accepted, accepted, cancelled, cancelled, reje…
```

Objectives

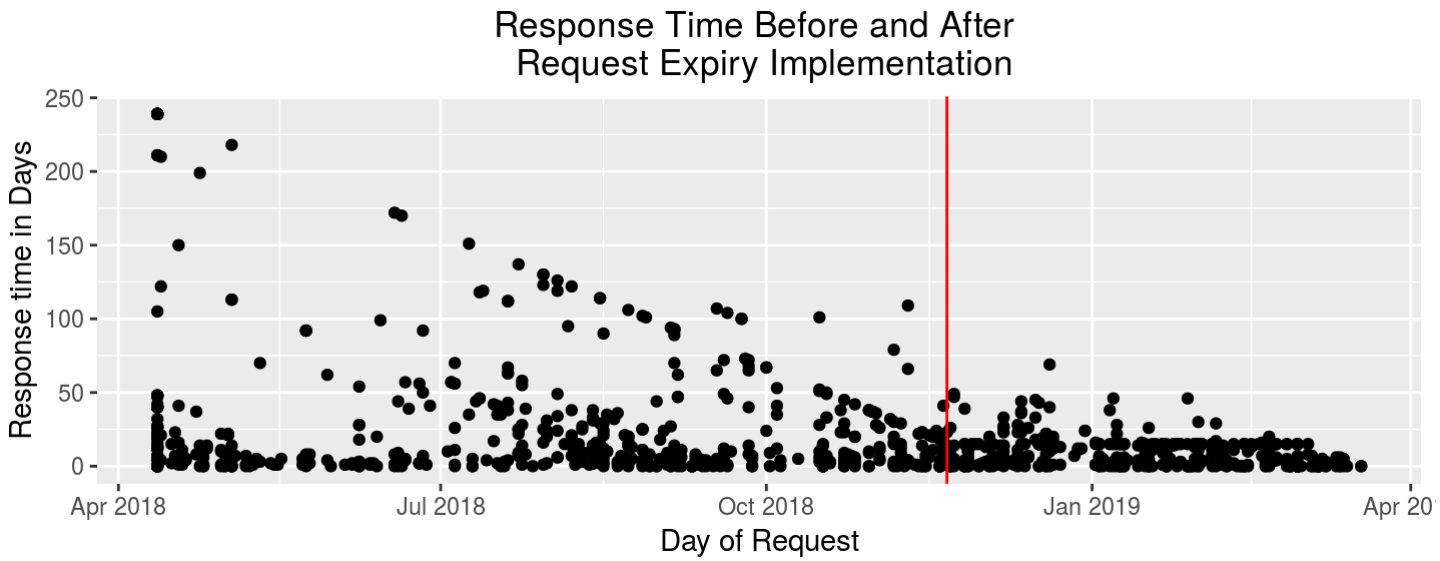
On November 21st, 2018, Riipen introduced a “request expiry” feature where users must respond to requests within 14 days or they expire.

- Did the implementation of this feature affect volume of requests?
- How did the acceptance rate of requests change after the new feature was introduced?
- What about the response time?

Data Summary: Cleaning the Data

- Created two new variables: response_time and featureimplemented
- Filtered out currently pending requests
- Filtered out August 30th

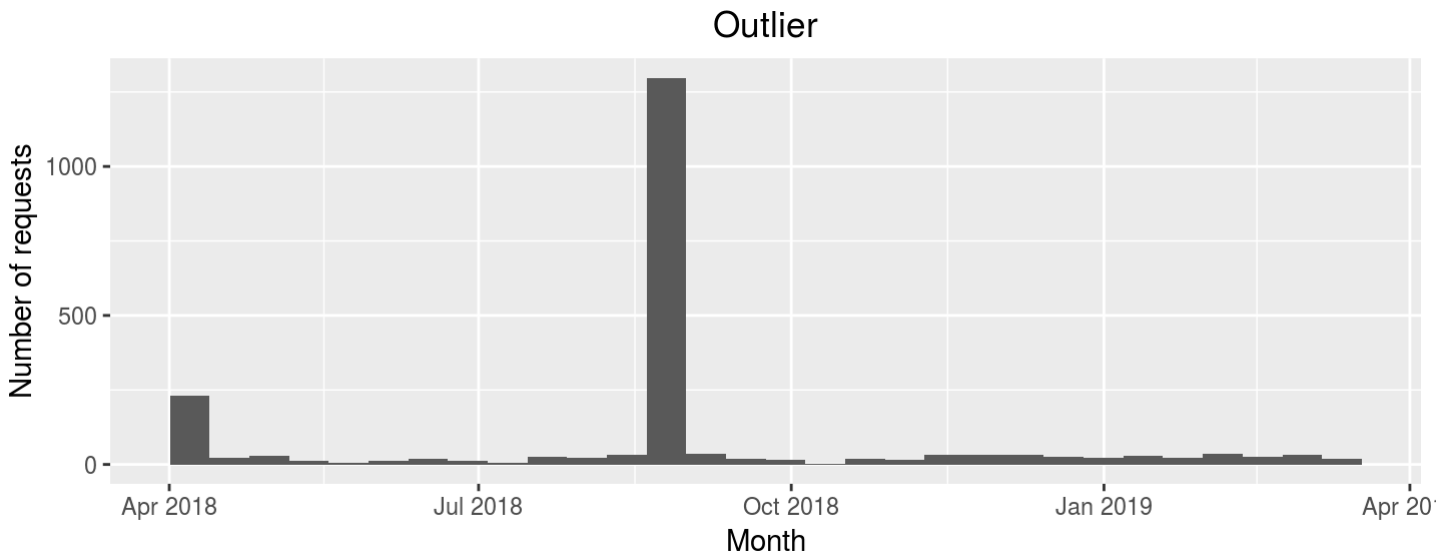
Overview of Data



Data Summary: Why Did We Remove August 30th?

- More than half the observations in the dataset were on august 30th
- All of them were responded to in 0 days
- None of them had ids
- We concluded that a server glitch must have occured on that day

Why we Removed August 30th



How August 30th Skewed the Dataset

```
## # A tibble: 2 x 2
##   `Day.of.Created.At == "30 August, 2018"` `number of requests`
##   <lgl>                                                  <int>
## 1 FALSE                                                    880
## 2 TRUE                                                    1269
```

Overall Acceptance rate, Before and After New Feature

- Overall: 43.6%
- Before implementation of request-expiry: 38%
- After implementation: 52%

Mean and Median Request Response Time: Before and After Feature

Before Feature

- Mean response time of 25 days
- Median response time of 20 days

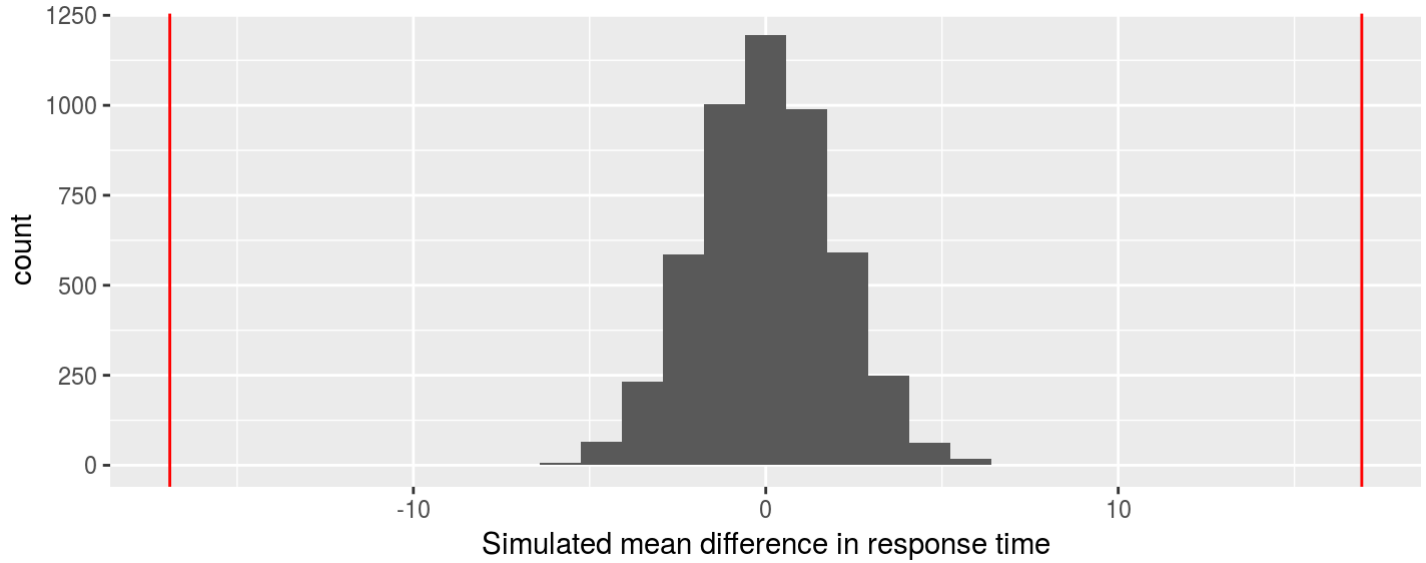
After Feature

- Mean response time of 8 days
- Median response time of 5 days

Statistical Methods

- Significance testing on response time before and after feature
  - H_0: Response time after - response time before = 0
  - H_A: Response time after - response time before ≠ 0
- Significance testing on request acceptance rate before and after feature
  - H_0: Acceptance rate after - Acceptance rate before = 0
  - H_A: Acceptance rate after - Acceptance rate before ≠ 0

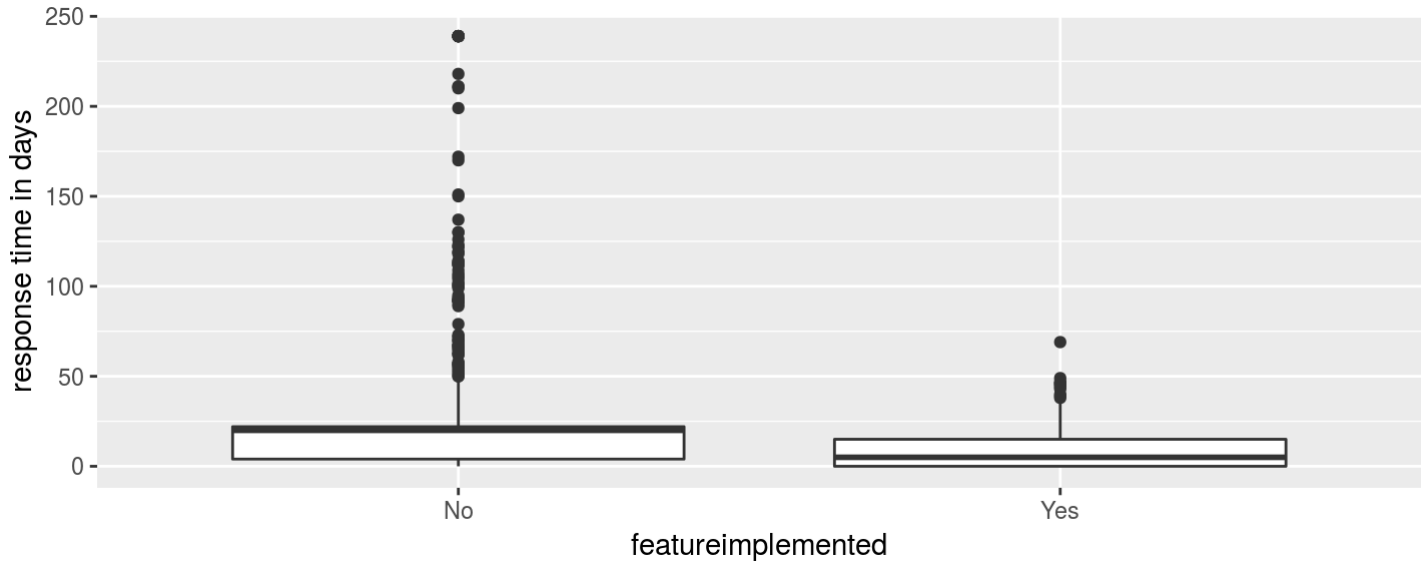
Mean Response Time Significance Test



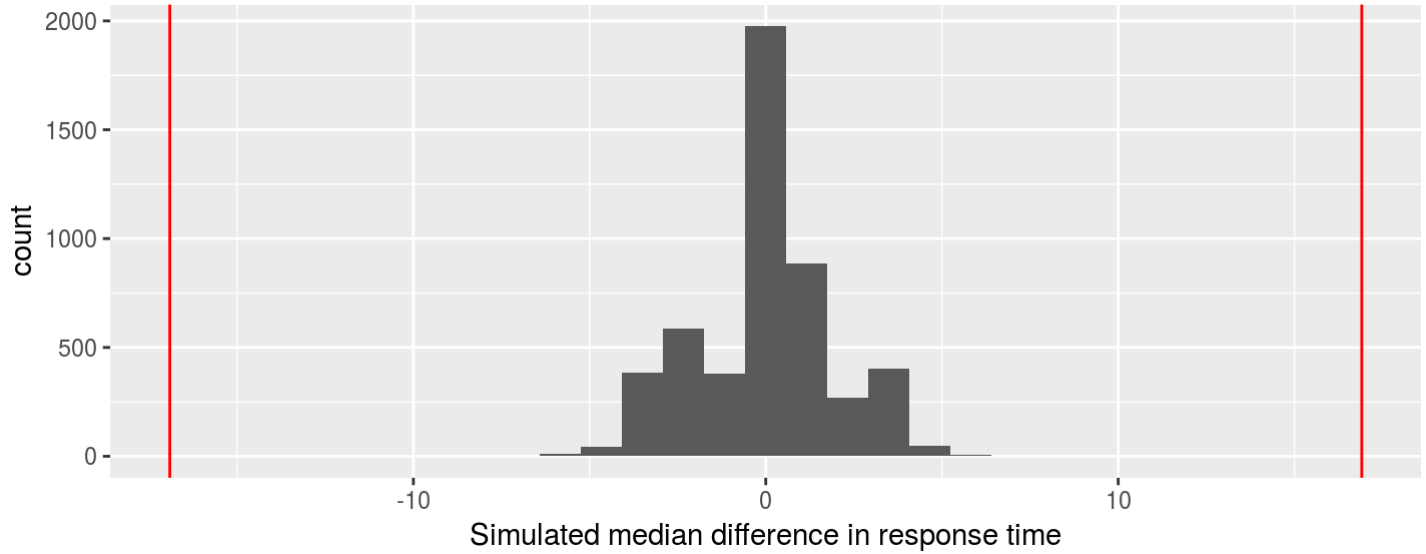
This is a histogram of all the simulated statistics, where our test statistic is represented by the red lines. The p-value is 0.00.

Boxplot of Response Times

```
requests_grouped %>%
  ggplot(mapping = aes(x = featureimplemented, y = response_time)) + geom_boxplot()
```

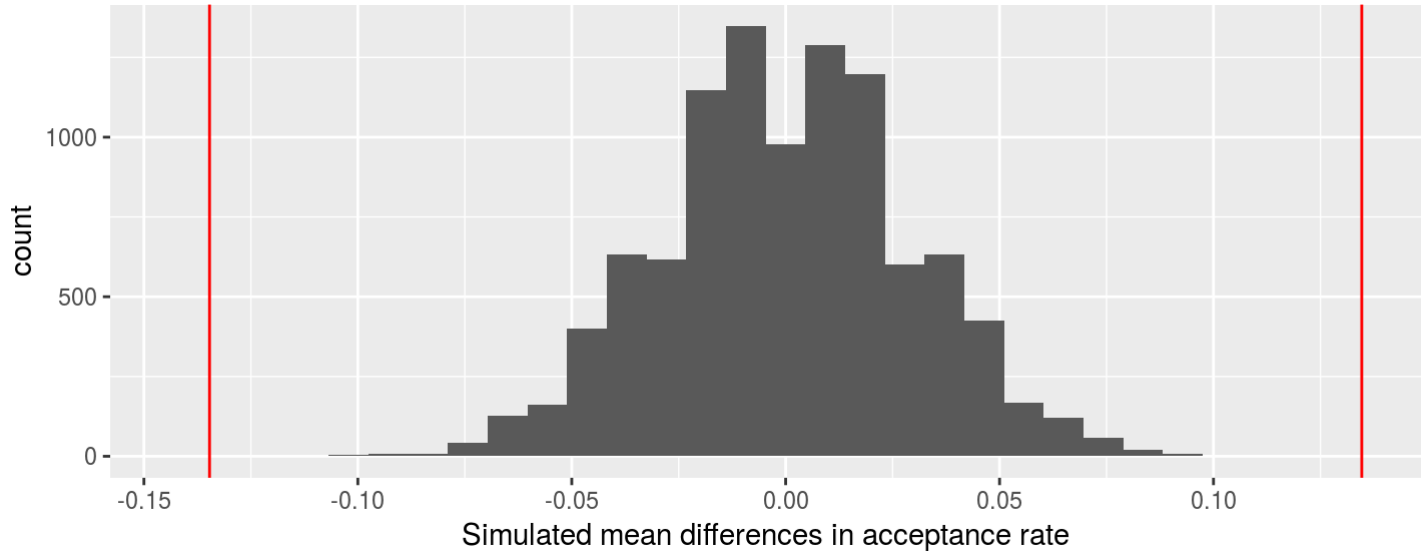


Significance Testing for Median Response Time



When using median instead of mean, the p-value is also 0.00.

Significance Testing for Mean Request Acceptance Rate



We obtained a p-value of 0.00.

Conclusion

- There is a strong evidence against the response time and request acceptance rate being the same after the implementation of the request expiry feature

Type I/II Errors

- Since we rejected the null-hypothesis, there is a possiblity of type I errors, I.E rejecting the null-hypothesis when we shouldn't have
- However, our p-values are so low that this is not really a concern
- No possibility for Type II errors because we did not uphold any null-hypotheses

Why Did We Obtain a P-value of 0?

- Theorectically, it is impossible to have p-value of 0
- Recall that we filtered out requests that was created on August 30th 2018, and only used 5000 simulations
- This could be the reason

Limitations

- No data for August 30th 2018
- 5000 simulations may or may not be enough

Acknowledgement

-We would like to thank Emily Masching, the Product Owner of Riipen who provided us to work with the most up-to-date data source.

Thank you for providing us this opportunity to learn more about statistics.