Fruit Fly Feeding Behaviour

Result and Analysis

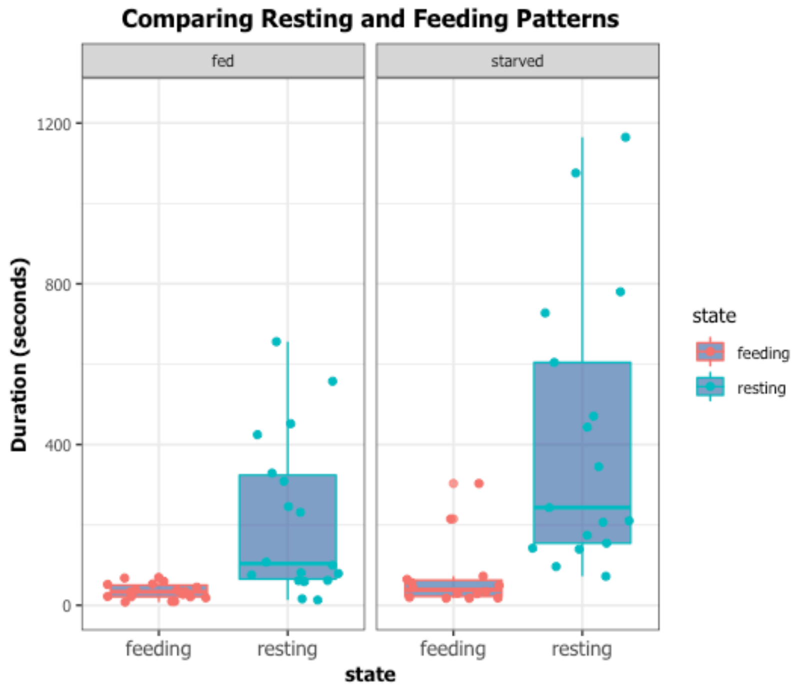
Yuji Jeong

yjeon009@uottawa.ca

March 11, 2019

TABLE OF CONTENTS

1. Boxplots: Comparing Resting and Feeding Patterns Between Fed and Starved Flies



The boxplots represent the five-number summary of the lengths of a given state intervals - either feeding or resting - in seconds. The five numbers are the first, second and third quartiles as well as the minimum and maximum of the data. The left compartment contains the boxplots of the three flies that were fed prior to recording the data. The right compartment contains the boxplots of the three other flies that were not given glucose water for a long period of time ("starved") prior to recording the data. See Appendix A for the boxplots of individual flies.

The boxplots above show that the medians duration of feeding intervals are not significantly different between the fed and starved flies. However, it can be observed that the resting intervals for the starved flies are significantly longer than that of the fed flies. Summary of the data can be obtained as the following:

	FED FLIES	**STARVED FLIES**
FEEDING INTERVAL	Min. 1st Qu. Median Mean 3rd Qu. Max. 8.06 21.18 33.59 34.80 50.03 69.43	Min. 1st Qu. Median Mean 3rd Qu. Max. 17.55 22.17 38.49 69.24 62.40 303.04
RESTING INTERVAL	Min. 1st Qu. Median Mean 3rd Qu. Max. 13.23 65.35 103.59 214.40 323.87 656.11	Min. 1st Qu. Median Mean 3rd Qu. Max. 72.21 154.43 243.15 414.73 604.31 1164.76

When feeding, the fed flies spend a median of 33.59 seconds, whereas the starved flies spend a median of 38.49 seconds or less. When resting, the fed flies spend a median of 103.59 seconds, whereas the starved flies spend a median of 414.73 seconds.

Before analyzing the difference between the two conditions of the flies, a test for normality and a test for homogeneity are done.

2. Normality and Homogeneity of Variance

Normality

We wish to understand the statistical properties of the data to select a suitable test to study the two groups of flies. We set the significance level of the tests at $\alpha = 0.05$.
Shaprio-Wilk test is used to check the null hypothesis that the data's distribution is not different from the normal distribution. Running the test on R, we obtain the following outputs.

Feeding Intervals Data

```
> residuals(feed.aov) %>% shapiro.test()

Shapiro-Wilk normality test

data:  .
W = 0.68069, p-value = 4.528e-07
```

Resting Intervals Data

```
> residuals(rest.aov) %>% shapiro.test()

Shapiro-Wilk normality test

data:  .
W = 0.89102, p-value = 0.002288
```

For both sets of data, p-values are below the significance level 0.05 meaning, the probabilities that the distribution of the data being normal are unlikely. We reject the null hypothesis. Therefore, we cannot assume normality. We must note that normality test is sensitive to the size of the data and that small-sized data tend to easily pass the test. In this case, we cannot be certain of the distribution of the data until larger dataset is acquired and studied. In this analysis, we will assume that the datasets are **not** normal.

Homogeneity of Variance

We conduct *Levene*'s test for the null hypothesis that the variance across the two groups, fed and starved flies, is not different.

Feeding Intervals Data

```
> leveneTest(duration ~ condition, data = feeding)
Levene's Test for Homogeneity of Variance (center = median)

      Df F value Pr(>F)
group  1  2.6685 0.1128
      30
```

Resting Intervals Data

```
> leveneTest(duration ~ condition, data = resting)

Levene's Test for Homogeneity of Variance (center = median)

      Df F value Pr(>F)
group  1  1.9086 0.1764
      33
```

For both sets of data, the p-values are greater than the significance level 0.05 meaning, the probabilities of having homogeneous variance across the two factor groups are likely. We do not reject the null hypothesis and assume that the variance across the fed and starved fly groups are not different for this analysis.

3. One-Factor Analysis of Variance

The assumptions required for ANOVA are 1) the distribution of the data is normal and 2) the variance across the different factor groups are homogeneous. In the previous section, we showed that normality is not assumed but homogeneity is assumed. For this reason, we conduct the non-parametric alternative ANOVA test, known as *Kruskal-Wallis* rank sum test. We test the null hypothesis that the medians of the two factor groups are equal.

Feeding Intervals Data

```
> kruskal.test(duration ~ condition, data = feeding)

Kruskal-Wallis rank sum test

data:  duration by condition
Kruskal-Wallis chi-squared = 0.69841, df = 1, p-value = 0.4033
```

Resting Intervals Data


```
> kruskal.test(duration ~ condition, data = resting)
Kruskal-Wallis rank sum test

data:  duration by condition
Kruskal-Wallis chi-squared = 4.6024, df = 1, p-value = 0.0319
```

For the feeding interval dataset, the p-value is greater than the significance level 0.05. We do not reject the hypothesis that the median feeding duration of the fed and starved flies groups are equal. For the resting interval dataset, the p-value is smaller than 0.05. We reject the hypothesis that the median resting duration of the fed and starved flies group are equal. The result of the analysis confirms the assumptions made from the boxplots on page 3.

We can assume that there is a significant difference in the resting intervals between the two groups, however the difference seems negligible for the feeding intervals.

4. Sources of Error

Six flies were used to record the data. Three of those flies were "starved" and the other three were "fed" before conducting the experiment. Exposing each fly to one condition and later another, can be considered to obtain sets of data that are more scientifically comparable.

The definitions of "feeding" and "resting", and the time requirements give drastically different results. The assumptions associated with what is considered as a "true" feeding/resting interval also significantly affect the results. It is recommended to set a time requirement for *only* feeding interval so that the time spent other than feeding would, automatically, be considered as resting state. As demonstrated in the programming script, setting the time requirements for both feeding and resting intervals requires a more complicated algorithm to prepare the data for analysis. If the algorithm can be simplified, another source of error can be potentially avoided.

Another source of error is the lack of data. After cleaning and preparing the data, 67 observations are obtained. Try

```
> length(compare_boxes$duration)
```

For better representation of the distribution and variance of the data, it is recommended to collect a larger set of data.

Lastly, it should be noted that the interpretation of the tests are affected by the level of significance. For example, in the section One-Factor Analysis of Variance, the p-value for the resting intervals data is 0.03193. If we had set the level at $\alpha = 0.01$, then the null hypothesis would be accepted and we would conclude that there is no difference between the groups when resting. Significance level represents the probability of Type I error, which is the probability of accepting the null hypothesis when in fact the alternative hypothesis is true. It is desirable to avoid the error and so testing with $\alpha = 0.01$ may be preferred over $\alpha = 0.05$, depending on the researcher.

5. Appendix

Boxplots by individual flies