Formal Report of Elution experiment

Kosuke Sasaki, Carolyn Wright, Guangze Yu

11/12/2021

I. Introduction

This client is looking to understand the relationship between elution time and the amount of sperm extracted from cotton swabs. In other words, she is looking to test for a significant difference in the relative percentage of collected sperm between the four groups with differing lengths of time of elution. The client already has a dataset that was collected through the procedure of the experiment. The data contains the relative percentages of each replicate, by group. The client's goal is to test for a significant difference in the relative percentage of collected sperm between the four groups. Our procedure divided into two parts: Checking for the assumption and checking ANOVA test.

II. Experiment setting

The procedures of the experiment are as follows:

1. Sample: Take samples with 3 swabs and cut them into 4 pieces equally
    - Note: the client only used 10 of the 12 total pieces for each grouping
2. Elution: The process of using a solvent to extract an adsorbed substance from a solid adsorbing medium.
    - Substance: sperm
    - Adsorbing medium: cotton swab
3. Centrifugation: Separate substance from absorbing medium by applying centrifugal force for 5 minutes.
    - Note: this time is held constant across all groups
4. Count: Count the number of substances in the pellet.

The setting of the experiment is as follows:

1. 2 elutions per each ¼ swab
2. Grouped by time of elution
    - Group 1: (10 replicates)
        - First elution: 1 min
        - Second elution: 1hr 59min
    - Group 2: (10 replicates)
        - First elution: 30 min
        - Second elution: 1hr 30min
    - Group 3: (10 replicates)

  - First elution: 1hr
  - Second elution: 1hr
- Group 4: (10 replicates)
  - First elution: 2hr
  - Second elution: 0hr

III. Major test method

Our experiment data structure is already shown above. Based on the character of our data, we decide to use log odds to eliminate the influence of different sample sizes for different replicates. Log odds are defined as the chances of success divided by the chances of failure. In our case, log odds are defined by the probability of the number of sperm of first elution divided by the probability of the number of sperm of second elution. The reason is that using log odds is easily updated with new data. The advantage for log odds is that it is easy to visualize variables without changing the characters of variables and also make variables distributed more normal. It will avoid violating assumption of any further test.

Then, we use one-way ANOVA (analysis of variance), also known as one-factor ANOVA, which is an extension of independent two-samples t-test for comparing means in a situation where there are more than two groups.

There are two major assumptions for one-way ANOVA. The data are normally distributed and the variance across groups is homogeneous.

To check the normality, we check the density plot for each group. The density plot is a representation of the distribution of a numeric variable. It uses a kernel density estimate to show the probability density function of the variable. In our case, we use density plots for each group whether they have normal distribution or not. Then, we plot the QQ (quantile-quantile) plot for each group. QQ plot is a graphical technique for determining if two data sets come from populations with a normal distribution. To be precise, we test for Shapiro Test. Shapiro Test is a test of normality in frequentist statistics. In our setting, we choose the alpha value as 0.05.

Then, we check whether the variance across groups is homogeneous. The residuals versus fits plot can be used to check the homogeneity of variances. Also, we test Levene's test, which is less sensitive to departures from the normal distribution in case our data is not a normal distribution.

After checking assumptions, we test for one-way ANOVA to test whether there is any siginificiant different between groups. The result and discussion are shown below.

IV. Result & Discussion

1. Data Processing

The head five line of original data is shown below. The shown data only contain part of information of group 1. The full data contain full information of four different groups with each 10 replicates.

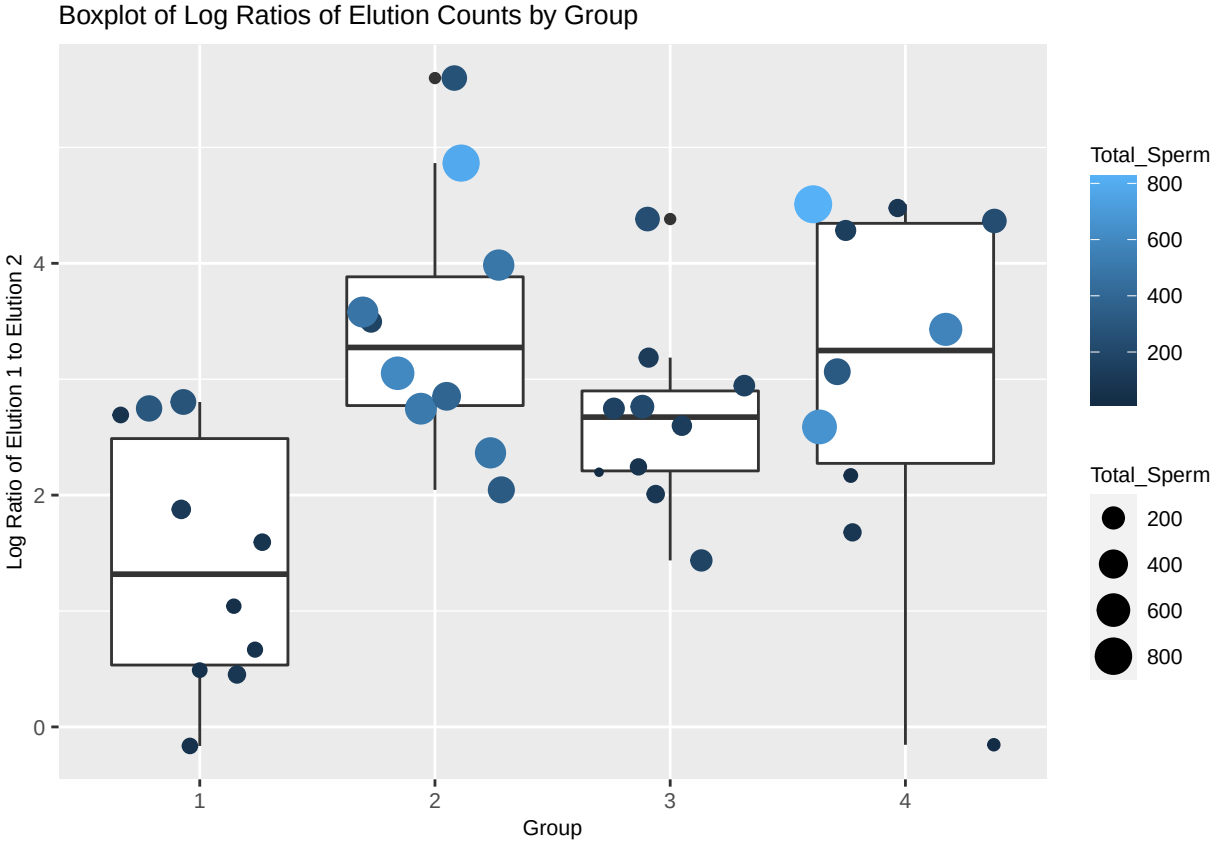
Sample	Sperm.Count	Total.Sperm.for.Replicate	Relative.%	group
R#1E#1	98	113	0.8672566	1
R#1E#2	15	NA	0.1327434	1
R#2E#1	37	56	0.6607143	1
R#2E#2	19	NA	0.3392857	1
R#3E#1	34	46	0.7391304	1

After cleaning and processing data, we clean the original data into following format and also calculate the log odds for each replicates.

Replicate	group	Sperm.Count_1	Sperm.Count_2	Relative.Pct_1	Relative.Pct_2	log_ratio	Total_Sperm
1	1	98	15	0.8672566	0.1327434	1.8769173	113
2	1	37	19	0.6607143	0.3392857	0.6664789	56
3	1	34	12	0.7391304	0.2608696	1.0414539	46
4	1	28	33	0.4590164	0.5409836	-0.1643031	61
5	1	55	35	0.6111111	0.3888889	0.4519851	90

We visualize our log odds ratio of each replicate with boxplots. A box plot or boxplot is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending from the boxes indicating variability outside the upper and lower quartiles. In our case, the value of the x-axis represents the group belonging to this replicate. The value of the y-axis represents the log-odds value of Elution 1 to Elution 2 for this replicate. The darker blue and smaller spots represent that this replicate has a small total sperm number from two elution times. The lighter and bigger spots represent that this replicate has a large total sperm number from two elution times.

From the boxplots, we can investigate that group 1 has the lowest log-odds median ratio among the four groups. There is one outlier for both group 2 and group 3.

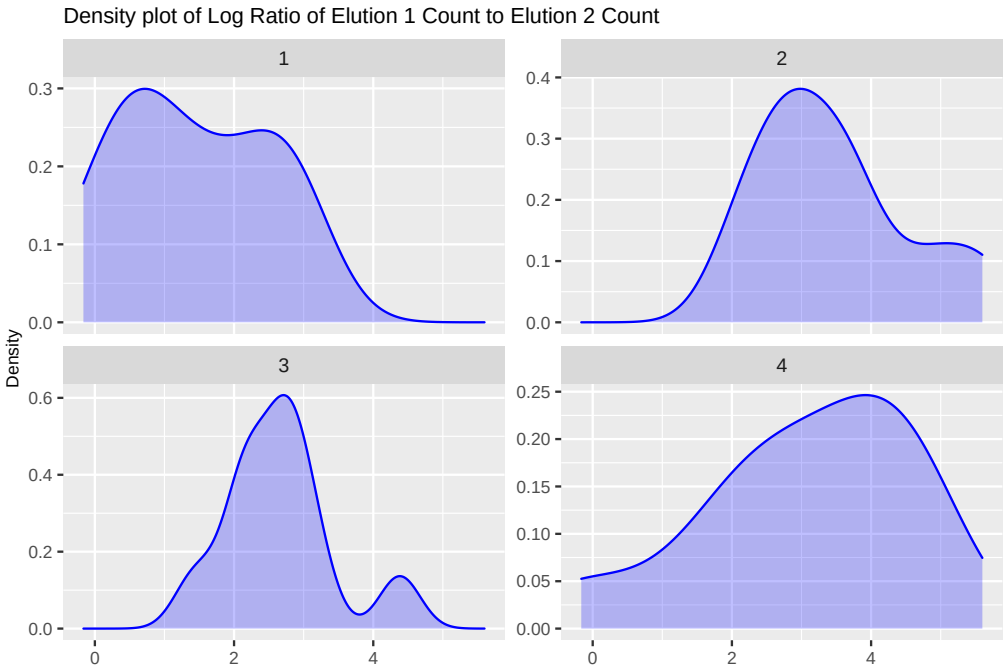


2. Check assumption

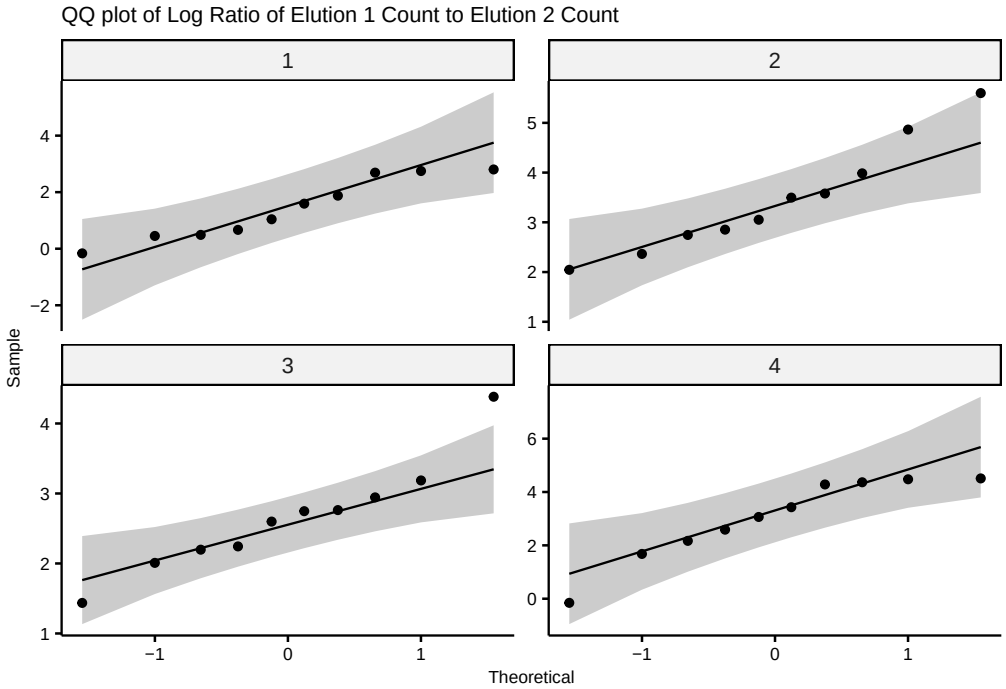
a. Normality

As the description in part III., our steps are density plots, the QQ plots and Shapiro Test.

Density plots for four groups are shown below. From density plots, we can investigate that four groups have approximate bell shape.



QQ plots for four groups are shown below. From QQ plots, we can investigate that four groups approximate within the estimated range.



To be more precise, we test for Shapiro Test. We extracted all p-value from four groups and summary into below chart. The null hyphothesis for Shapiro Test is that the population is normally distributed. Thus, if the p value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed. On the other hand, if the p value is greater than the chosen alpha level, then the null hypothesis (that the data came from a normally distributed population) can not be

rejected. In our case, our alpha value is 0.05. All groups' p-value are larger than 0.05. Thus, we accept the null hypothesis. We can conclude that all four groups are normal distributed.

Statistic	P_Value	Group
0.9107267	0.2860200	Group 1
0.9433447	0.5907949	Group 2
0.9416273	0.5712591	Group 3
0.8890707	0.1655346	Group 4

b. Homogeneous

As the description in Part III, we use Levene's test to test the variance across groups. The null hypothesis is that the population variances are equal. In our case, the p-value is larger than 0.05. Thus, we accept the null hypothesis. We can concluse the pupulation variances are equal.

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  1.4241 0.2517
##       36
```

3. ANOVA analysis

After checking the assumptions of ANOVA analysis, we can test ANOVA over four groups. The null hypothesis is that there is no significiant difference between groups. We attached the result of ANOVA below. Since the p-value is smaller than the alpha value, we can conclude that there must be at least one group is siginificiant different than other groups. Then, we test for Tukey multiple pairwise-comparsions.

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## group        3  23.20   7.732     5.8 0.00243 **
## Residuals   36  47.99   1.333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For Tukey multiple pairwise-comparsions, we attched the result as below chart. As the p-value is less than the significance level 0.05, we can conclude that there are significant differences between the group 2-1 and the group 4-1. We can also valid from 95% confidence interval. The confident interval difference between group 2 and group 1 doesn't contain 0. The confident intervaldifference between group 4 and group 1 doesn't contain 0. Also, we test for the normality of ANOVA residual. The p-value is larger than the alpha value. We can conclude that the residual is normal distributed.

	Diff	Lower Bound	Upper Bound	p-value
Group 2-1	2.04	0.65	3.43	0.00
Group 3-1	1.23	-0.16	2.62	0.10
Group 4-1	1.62	0.23	3.01	0.02
Group 3-2	-0.81	-2.20	0.58	0.41
Group 4-2	-0.42	-1.81	0.97	0.85
Group 4-3	0.39	-1.00	1.78	0.87

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.96515, p-value = 0.2501
```

V. Conclusion

Based on our anlysis of ANOVA, if the p-value is smaller than our alpha value,we can conclude that there is significant difference between groups. Then, we conclude that there is significiant difference between the ratio of group 2 and group 1. Also, there is significant difference between the ratio of group 4 and group 1.