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. Significant Difference Testing Method

Our experiment data structure is already shown above. Based on the character of our data, we decide to use log ratio of the number of counts of elution 1 to elution 2 instead of ratio to compare the number of counts of elution 1 and 2 between the groups. Log transformation can make it easy to see the difference of the mean ratios between the groups as well as reduce the skewness of the distribution of the ratio, which makes the distribution closer to the normal distribution.

Then, we used 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 checked 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 used the Shapiro Test. The Shapiro Test is a test of normality in frequentist statistics. In our setting, we choose the alpha value as 0.05.

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

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

IV. Result & Discussion

1. Data Processing

The first five lines of original data is shown below. The shown data only contains part of the information of group 1. The full data contains all of the information for the four different groups with each of the 10 replicates.

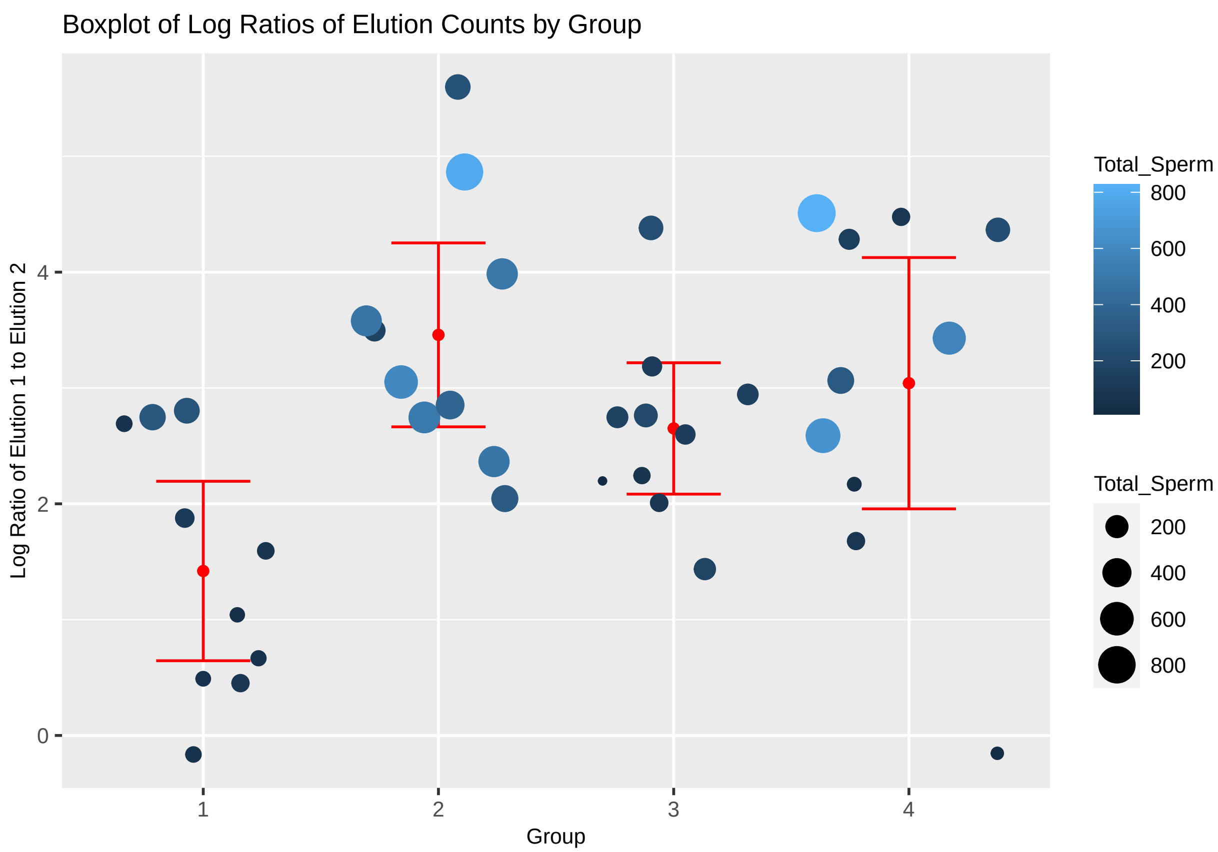
Sample	Sperm.Count	Total.Sperm.for.Replicate	Relative.%	group
R#1E#1	98	113	0.8672566	1
R#1E#2	15	NA	0.1327434	1
R#2E#1	37	56	0.6607143	1
R#2E#2	19	NA	0.3392857	1
R#3E#1	34	46	0.7391304	1

Next, we transformed the original data into the following format and calculated the log ratio for each replicate.

Replicate	group	Sperm.Count_1	Sperm.Count_2	Relative.Pct_1	Relative.Pct_2	log_ratio	ratio	Total_Sperm
1	1	98	15	0.8672566	0.1327434	1.8769173	6.5333333	113
2	1	37	19	0.6607143	0.3392857	0.6664789	1.9473684	56
3	1	34	12	0.7391304	0.2608696	1.0414539	2.8333333	46
4	1	28	33	0.4590164	0.5409836	-0.1643031	0.8484848	61
5	1	55	35	0.6111111	0.3888889	0.4519851	1.5714286	90

We visualized log ratio of the number of counts of elution 1 to elution 2 in each replicate as a blue dot with error bar shown as a red line. The error bar represents 95% confidence interval extending around the mean of log ratios in each group which is shown as a red dot. The darker blue and smaller point represents smaller total sperm number from two elution times. The lighter and bigger point represents a larger total sperm number from two elution times.

Based on the graph, group 1 seems to have a mean log ratio lower than that of the other three groups.

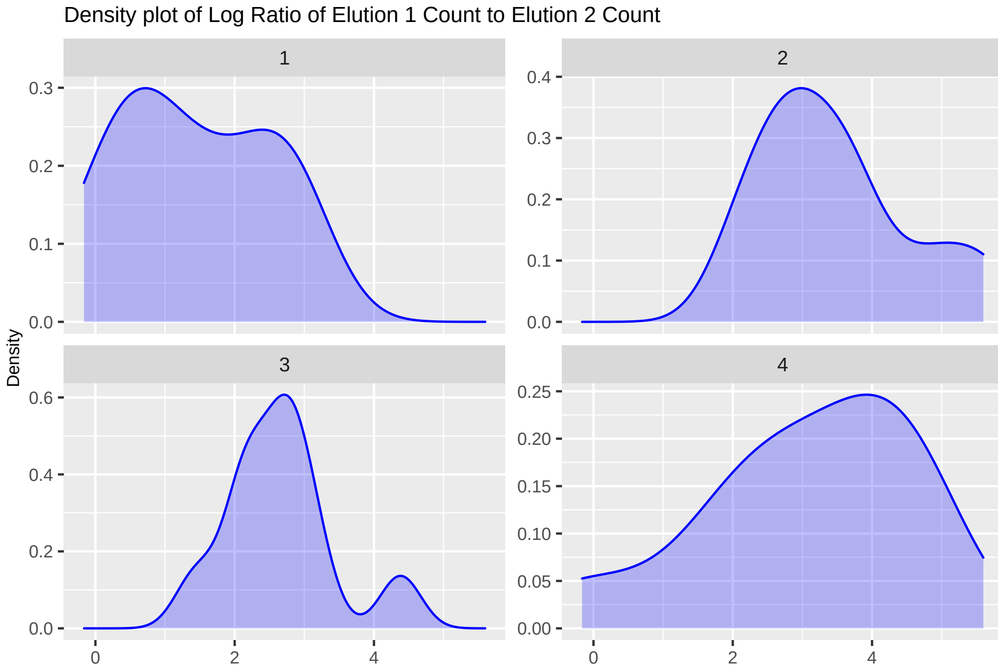


2. Check assumption

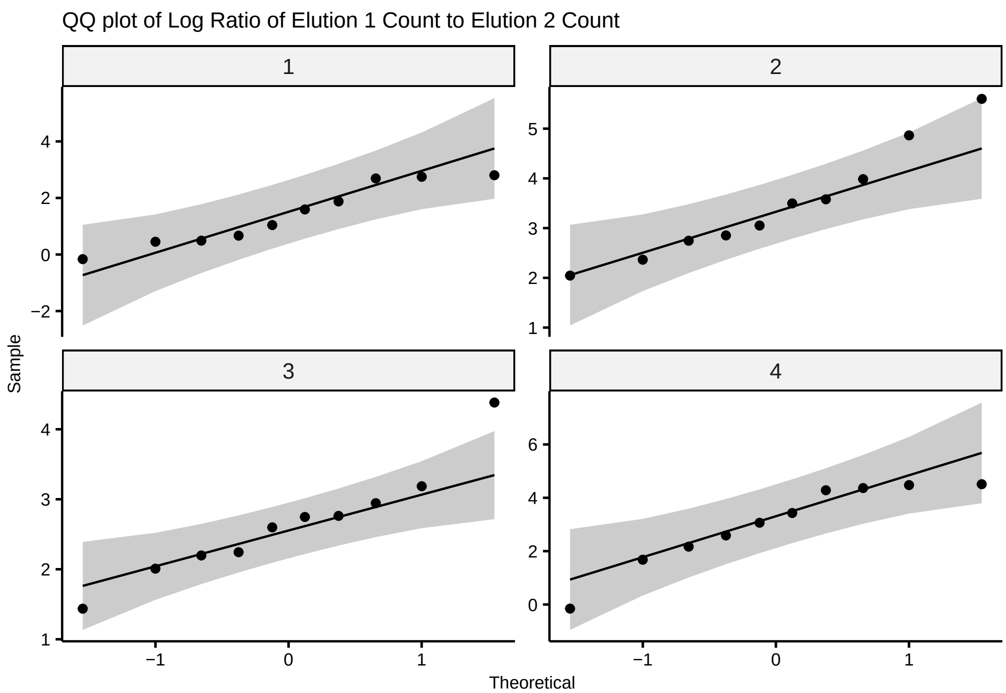
a. Normality

As described in part III., we used the following processes to validate our normality assumptions: density plots, the QQ plots and Shapiro Test.

Density plots for four groups are shown below. From density plots, we can see that four groups are all approximately bell shaped.



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



To be more precise, we used the Shapiro Test to check this assumption numerically. 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 normally distributed.

Shapiro Test		
Group	Test Statistic	P-Value
Group 1	0.9107267	0.2860200
Group 2	0.9433447	0.5907949
Group 3	0.9416273	0.5712591
Group 4	0.8890707	.1655346

b. Homogeneous

As the described in Part III, we used 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	
F-value	P-Value
1.4241	.2517

3. ANOVA analysis

After checking the assumptions of the ANOVA analysis, we ran an ANOVA test scross the 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 significiantly different than other groups. To validate this test, we also tested the ANOVA residuals for the normalityl. The p-value is larger than the alpha value. We can conclude that the residual is normal distributed. Howvever, this test does not specify which groups contain the differences. To identify the specific groups that differ we used Tukey multiple pairwise-comparsions.

ANOVA Results

ANOVA Results					
	Df	Sum Sq	Mean Sq	F-Value	P-Value
group	3	23.20	7.732	5.8	.00243
Residuals	36	47.99	1.333		

Validation of ANOVA - Shapiro Test of residuals

Shapiro-Wilk normality test	
Test Statistic	P-Value
.96515	.2501

The results for the Tukey multiple pariwise-comparisons, are listed below. As the p-value is less than the significance level 0.05, we can conclude that there are significant differences between group 2 and group 1 as well as between group 4 and group 1. We can also see this from looking at the 95% confidence interval. The confidence interval relating to the comparison between group 2 and group 1 does not contain 0. Additionally, the confidence interval relating to the comparison between group 4 and group 1 doesn't contain 0.

	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

4. Sample size calculation

We also calculated sample size needed to detect any difference in the mean log ratio of the counts between group 1 and any other group with time of elution 1 longer than 30 minutes. Assuming the true mean of group 1 is 1.42, the true mean of any group with time of elution 1 larger than 30 minutes is the mean of those of group 2, 3 and 4, or 3.05, and the pooled standard deviation is 1.15, the needed sample size is 14 for each group at the power of 0.8 and 0.008 significance level.

V. Conclusion

Based on our analysis of ANOVA, if the p-value is smaller than our alpha value,we can conclude that there is significant difference between groups. Then, based on the Tukey multiple pairwise-comparisons, we can conclude that there is a significant difference between the ratio of group 2 and group 1 as well as between group 4 and group 1. Whereas we cannot find any significant difference between group 3 and group1. This could be due to sampling error because the ratio of group 3 should be greater than group 2 theoretically. To detect the difference between the group 3 and group 1, 14 replicates for each group would be needed.