1 Introduction

Cigarette smoking has several healthy effects. For example, heart disease, stroke, and lung cancer. The worse thing it can causes deaths. The pregnancy women that smoking also has many effects for the unborn baby, especially in the lung and brain. Moreover, it also has effects to the birth weight. This project focuses on multiple pairwise comparisons, which are carried out using two-sample t-tests. Furthermore, we attempt to adjust the p-value using the Bonferroni and Tukey methods. The dataset of this project is a collection of values of mother and their babies data. It contain 1236 sample size with 23 independent variables. The independent variables in this dataset are infant survival, birth weight, date of birth, sex, mother's ethnicity, age, education level, height, weight, and smoking status. The dataset divide into five smoking mother history which are, never, smokes now, smoke until current pregnancy, smoke once but not now, and unknown. The purpose of this assessment is to understand the trends in maternal smoking and babies weight. Furthermore, the confidence interval for Tukey's method is also analyse. The second chapter describes the dataset and the goals of this project.

In the third section, the statistical methodologies used for this project, such as statistical hypothesis testing, the one-way ANOVA test, the two-sample t-test, and the multiple testing problems, are all explained. Several assumptions must be met first before calculating the ANOVA. Assumption of ANOVA, independence, normality, and equality of variance must be satisfied. For normality check, the visualization method, normal quantile-quantile (Q-Q) plot, is used and explained in detail in this section. Boxplot is used to check the equality of variance. When the assumption of ANOVA is met, ANOVA can be performed. In the fourth section, interprets the test results of one-way ANOVA and two-sample t-tests, along with Bonferonni and Tukey's methods. The graphs are drawn based on the methods introduced in the previous section for a more detailed analysis. The last chapter is about summaries the main results and the possibility for further analysis that can be done using this dataset.

2 Problem statement

2.1 Description of the dataset

This report is an analysis of the dataset provided by the lecturers of Introductory Case Studies at TU Dortmund University in the summer semester of 2023. The data have been collected by the lecturers via the web Berkeley Statistics (Statlabs, 2002). The dataset provided for this report covers the 23 independent variables, including the baby's weight and smoking mother category, divided into five categories. The baby's birth weight is shown in ounces. There are ten missing values in the baby's birth weight column. Imputing missing values is necessary to analyze the data (Hastie and Tibshirani, 2008).

2.2 Project goals

The object of this project is to compare the birth weight of five smoking mother categories. Moreover, another goal is to check whether there are any differences in the birth weight between the given five categories of data. ANOVA is conducted to check that all the smoking mother categories have equal mean. Before performing the ANOVA test, three assumptions about ANOVA must be met. Visualized plots are used to check normality and homogeneity of variance. Another object is to compare pairs of smoking mother categories. If the ANOVA test rejects the hypothesis that all categories have the same mean, then a post hoc analysis should be performed to determine which groups are different. A pairwise t-test is conducted to check whether there is a difference between each two groups. A correction is necessary in this case, as multiple comparisons can weaken the power of the test. Another important object is to compare the values before correction and after correction with the Bonferroni method and also with the Tukey procedure.

3 Statistical methods

In this section, several statistical methods used for analysing the dataset in the later section are described. All tables and plots are created with the statistical software R in version 4.0.5 (R Development Core Team, 2020) using the statistical methods in this chapter. Used packages are readr, lessR, dplyr (Grolemund and Wickham, 2017, p. 3,43,125), xtable (Dahl and Scott, 2019), ggplot2 (Grolemund and Wickham, 2017), ggpubr (Kassambara, 2022) and multcompView (Graves and PeterPiepho, 2019).

3.1 Statistical hypothesis test

The statistical hypothesis test is the statistical method of performing inference that examines hypotheses about parameters using samples extracted from the population (Roussas, 2003). A statistical hypothesis is denoted by H followed by the assertion that defines the hypothesis. The judgment of the assertion is based on sample data and clearly shows how likely or unlikely the assertion is to be correct or incorrect. A hypothesis can be simple or composite, depending on the parameter values. To test statistical hypotheses, two opposing hypotheses, the null hypothesis denoted by H_0 and the alternative hypothesis denoted by H_1, must be established. They are mutually exclusive. If the test is conducted under the assumption that the null hypothesis is correct and the null hypothesis is found to be false, the null hypothesis is rejected and the alternative hypothesis is accepted. Hypothesis tests are always likely to have errors. There are type I error(α-error) and type II error(α-error) (Selvamuthu and Das, 2018). Type I error occurs when the null hypothesis is rejected eventhough it is actually correct, and type II error occurs when the null hypothesis is accepted as correct eventhough it is actually false. In hypothesis testing, the type I error should be controlled and have an upper bound. The maximum allowable probability of a type I error is called the significance level(α). The significance level is usually set at 1% ($\alpha = 0.01$) or 5% ($\alpha = 0.05$).

The confidence level is the probability of deciding that null hypothesis is right when it is actually correct. The confidence level is equivalent to $1 - \alpha$. When the distribution of the test statistic is considered under the assumption of a true H_0, but the test itself is not necessarily conducted under the assumption that H_0 is true, the p-value states the probability of getting a value greater than the actual observation test statistic. If the p-value is less than the significance value, the null hypothesis is rejected. On the contrary, if the p-value is greater than the significance value, it fails to reject the null hypothesis. There are one-tailed test and two-tailed test. A one-tailed test is a test for the hypothesis, where the parameter is greater than or less than some value, while a two-tailed test is a test for the hypothesis, where the parameter is equal to some value.

3.2 Visualization methods

Q-Q plot

Normal quantile-quantile (Q-Q) plot is is used to check normality of sample distributions (Jahans, 2019). Q-Q plot compares the sample of data with standard normal distribution. When the dataset is arranged with size $y_1 < y_2 < .. < y_n$ and standard normal distribution divided into $n + 1$ in equal areas, each sample data is matched with the standard normal random variable $x_1 < x_2 < .. < x_n$. The smallest value in sample y_1 is matched with the smallest value expected from the standard normal distribution x_1. It also applies for the larger values in sample data, then they are plotted with the pair (x_i, y_i). The approximately straight line is created in the Q-Q plot from a sample value as a result of it being normally distributed (Dodge, 2010).

3.3 Analysis of variance

Analysis of variance or known as ANOVA is a statistical technique that is used to compare numerous groups (Davey and Doncaster, 2007). In this report, we use a one-way ANOVA test to compare the means of several populations. Several assumptions which are normality, independence of observations, and equality of variances should be fulfilled before ANOVA. The null hypothesis H_0 of ANOVA is the means between groups are equal $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ where k is the number of groups (Selvamuthu and Das, 2018, p.178-179). And the alternative hypothesis H_1 is not all means are equal. $H_1 : \mu_i \neq \mu_j$ for at least one pair i, j with $i \neq j$. That is, at least one group has different mean from the other groups. During the ANOVA process, F-test statistic is performed. The F value is the ratio between sample means to variance within the samples. Under the null hypothesis F distribution with degrees of freedom $(k - 1, n - k)$ where n is the number of all data. Table 1 shows ANOVA table (Verzani, 2014) with the calculations.

Df column shows the degrees of freedom. Sum sq is sum of squares and it represents the measures of variation from the mean value. The total sum of squares equals the

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	k-1	SSTr	MSTr	F	p-value
Residuals	k-1	SSE	MSE		

Table 1: ANOVA Table

treatment sum of squares plus the error sum of squares. The treatment sum of squares is the sum of the values obtained by subtracting the total mean from the sample mean of each population, and then multiplying the sample size. The formula of sum of squares treatment ($SSTr$) is $SSTr = \sum_{i=1}^{k} n_i(\bar{x_i} - \bar{x})^2$. It is a number that indicates the difference between groups. This value increases if the sample means are far apart from each other. The error sum of squares is the sum of the squares of residuals. It is the squared difference from the sample mean within the population $SSE = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x_i})^2$. If the values in the population are spread out, the SSE gets larger. The treatment mean square is the value of the treatment sum of squares devided by the degrees of freedom $MSTr = SSTr/(k-1)$. It shows the average variability between groups. The mean square of the error is the value of the error sum of squares devided by the degrees of freedom $MSE = SSE/(n-k)$. It shows the average variability within groups. The ratio of the mean square between groups to the mean square within a group is called the test statistic F (Selvamuthu and Das, 2018, p.180), $F = MSTr/MSE$. If the difference between means between the group is bigger, the F value also gets larger. That is, a large F-value means that there is a difference in the means between groups, and the null hypothesis is rejected. As a result of the significance test, the F statistic follows the F distribution with degrees of freedom $k-1$ and $n-k$, $F \sim F_{k-1,n-k}$ and if the calculated p-value is less than the set significant level (α), the null hypothesis is rejected. If the overall F value is not significance, it means that there is no significant difference in the means, so there is no need for a post hoc analysis.

3.4 Multiple Comparisons

In the F test, the null hypothesis is rejected when the p-value is less than the significance level (α). It indicates that there is at least one group has difference in mean value. Post hoc analysis is used to distinguish the groups which have different means. T-test is a multiple comparison test that is used in post hoc analysis for ANOVA (Muth, 2014). T-test operates in two groups of pairs to diagnose which group is different.

$K =_k C_2 = k(k-1)/2$ times for comparisons, where k is the number of groups are performed.

T-test

Pairwise t-test with pooled standard deviation is performed for multiple comparisons (Muth, 2014). To conduct t-test, several assumptions should be fulfilled. Data in each group must be independent and normally distributed and the variances of groups are equal.

$$t = \frac{\bar{X_1} - \bar{X_2}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\bar{X_1}$ and $\bar{X_2}$ are the means of the two groups to be compared and S_p is the pooled standard deviation. The equation for degree of freedom is $df = n - k$ where n is the number of sample and k is the number of groups. The following equation when more than two populations, pooled variance must include all samples.

$$S_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)}$$

where n_i is the sample size of group i and s_i is the sample variance of group i. It follows t distribution with degrees of freedom $n-k, T \sim t_{n-k}$ and if the calculated p-value from t-value is less than the set significance level, the null hypothesis is rejected.

In order to make the decision about the test, we have to compare the test statistic with the critical value of the t-test, and the critical value is calculated for our hypothesis assumption as $t_{1-\alpha/2,f}$ with α set to 0.05, and we reject the null hypothesis if the t statistic is greater than the critical value, that is, $|t| > t_{1-\alpha/2,f}$; similarly, we can reject the null hypothesis using the p-value approach if the associated p-value with the test statistic is less than the fixed significance level α (Ross, 2010, p.447-449). When we reject the null hypothesis even though it is true, we make a type I error. The significance level of the test (α) represents the probability of making a type I error. Another type of erroneous decision occurs when we not reject the null hypothesis despite the fact that the alternative hypothesis statement is true. This is referred to as a type II error, and the probability of committing such errors is represented by (β). The power of the test is represented by the complement of beta, which is $(1-\beta)$. The power is the probability of rejecting the null hypothesis when the null hypothesis is actually false. When we reduce the probability of making type I and type II errors, we have a better chance of getting an ideal test (Selvamuthu and Das, 2018, p. 147-148).

Bonferroni correction

One solution to resolve the multiple-testing problem is the Bonferroni correction. Bonferroni correlation adjusts the significance level divided by the number of comparisons of groups $\alpha * = \alpha / K$ (Jahans, 2019). Alternatively, the modified p-value is multiplied by the number of group comparisons $p - value* = p - value \times K$. Supposing that the calculated p-value is less than the modified significance level or the adjusted p-value is less than the significance level, the null hypothesis is then rejected as a result.

Tukey method

Tukey methods is one of the multiple comparison methods. Tukey procedure is a method that compares all pairs of means. Tukey also known as Tukey Honest. The equation to calculate Tukey is

$$T_\alpha = q_\alpha(a, f)\frac{MS_{Error}}{n}$$

Where a is a factor level, f is degree of freedom, n is number of treatment level, MS_{Error} is mean square error and $q_\alpha(a, f)$ is define as studentized range statistic corresponding to the level of significant value of α. For any pair j, $k = 1, 2, ..., p$ for which $\bar{y}_j \geq \bar{y}_k$, with formula $100(1 - \alpha)\%$. The equation for Tukey confidence interval for differences $\mu_j - \mu_k$ (Jahans, 2019) is

$$(\bar{y}_j - \bar{y}_k) - \frac{1}{\sqrt{2}} q_\alpha s \sqrt{\frac{1}{n_j} + \frac{1}{n_k}} < \mu_j - \mu_k < (\bar{y}_j - \bar{y}_k) + \frac{1}{\sqrt{2}} q_\alpha s \sqrt{\frac{1}{n_j} + \frac{1}{n_k}}$$

where s is define standard error, n is a the number of treatment levels. Tukey procedure is designing the significance level for individual and groups are same then α is compared with the p-value for each pairwise (Jahans, 2019).

4 Statistical analysis

This chapter covers the result in graphs from a calculation using mathematical methods in the previous chapter. It also covers the post hoc analysis with Bonferroni and Tukey's Honest method.

4.1 Frequency distribution

The dataset of this project is divided into five smoking mother category and the goal is to compare the Birth weight between all five histories. There are 10 missing values that are already mentioned in the previous chapter. Before doing the calculation methods, are necessary to replace the missing value. There are 10 missing value in column babies birth weight that need to be fulfilled manually. Moreover, the other missing value is replaced using mean. Since it suits the variables with normal distribution. Moreover, the "Unknown" category of smoking mother is not necessary to included in this analysis. Since the "Unknown" category not provide any information about mother smoking history. Figure 1 describes the distribution between each smoking mother category. The red dots in the boxplot represent the mean value of each history.

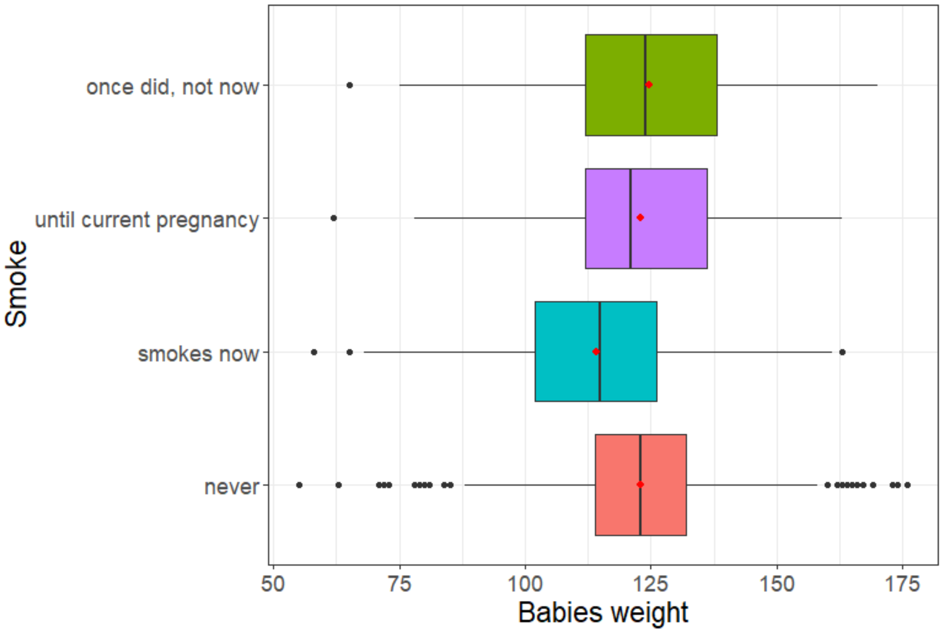


Figure 1: Boxplot of each categories

History	Count	Mean	Var	IQR
Once did, not now	101	124.58	341.67	26
Until current pregnancy	99	122.95	304.50	24
Smoke now	482	114.12	322.41	24
Never	544	122.83	288.99	18

Table 2: Summary of each smoking mother categories

The smoking mother category table and the boxplot sorted in the same way for easier viewing. Each smoking mother category has different sample size and the mean values between smoking mother category are numerically unequal. However, it is not necessarily confirmed as the ANOVA test will further verify whether or not it will be statistically significant.

4.2 Assumptions of ANOVA

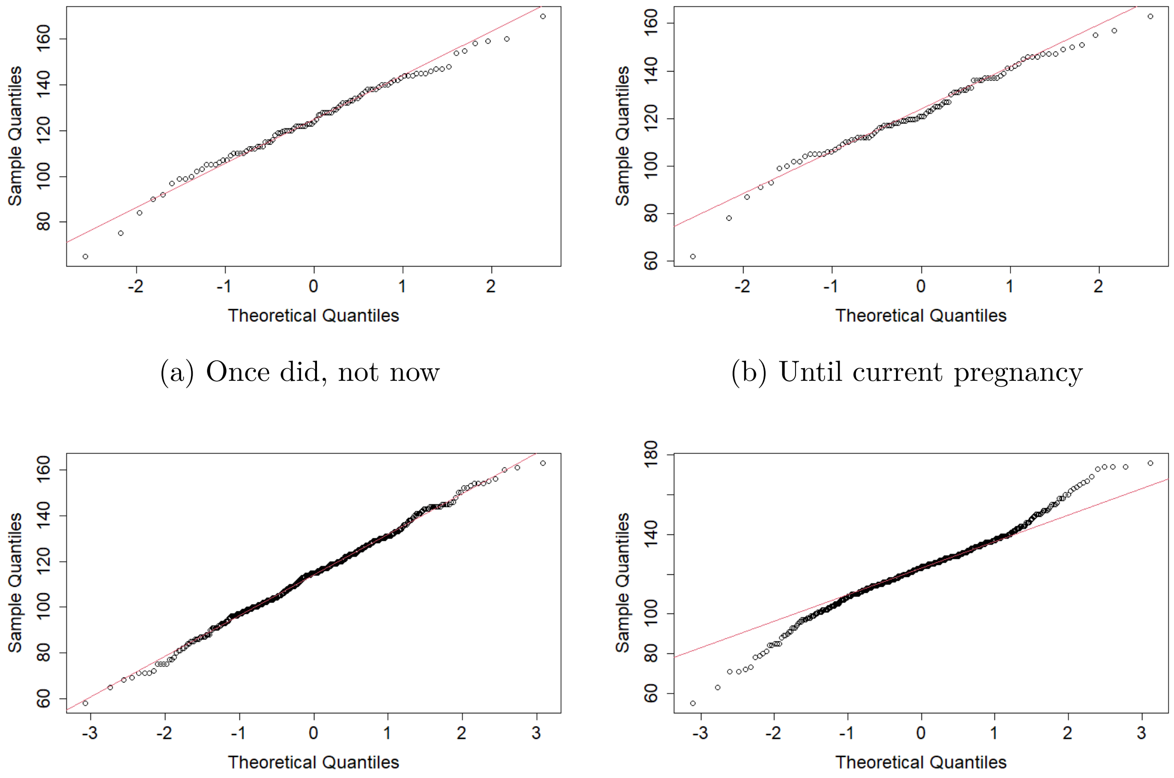
Before implementing the ANOVA, three assumptions need to be checked. Those assumptions are as follows: independence, normality, and equality of variance.

Independence

The dataset of this project is considered to be independent because it is taken randomly as a sample and each individual is only listed once.

Normality

Figure 2 shows Q-Q plots of 4 categories. The plots illustrate the comparison between sample distribution and theoretical distribution. The red line in the plot refers to the normal distribution. A normal distribution is achieved, and the normality is satisfied on the assumption when most of the data points lie in a straight line.



(a) Once did, not now

(b) Until current pregnancy

(c) Smoke now

(d) Never

Figure 2: Boxplots by subregion

Some extreme values can be seen in all groups. However, since the data points are generally distributed along the red line for all groups, we assume normality in this project and continue to test.

Equality of variance

The IQR length of each boxplot indicates the variance of babies weight of smoking mother category for each category. According to figure 1 and table 2 above, it is difficult to say that all categories have equal variance. However, the IQR value for the smoking mother category "Never" is slightly lower than for all other categories. However, Levene's test is perform and shows the p-value of the test is 0.106 which is greater than 0.05 then conclude that the variance among the four groups is equal.

4.3 One-way ANOVA

The dataset in this project consists of one dependent variable and one independent variable. For one dependent variable, one way ANOVA is used. The null hypothesis is a condition where means in all 4 categories are equal $H_0 : \mu_1 = ... = \mu_5$. Whereas the alternative hypothesis is that at least one of the means differs from others. The significance level in this project is set at 5% with the $\alpha = 0.05$. Table 4 shows the ANOVA table with the calculated statistics, which are used to test the hypothesis regarding to the means of population.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	3	23788.861	7929.620	25.770	3.630×10^{-16}
Residuals	1222	376004.545	307.69		

Table 3: ANOVA Table

In table 3, it can be seen that the p-value is much less than the significance level and hence we reject the null hypothesis that the means of all groups are equal and proceed with the post hoc analysis.

4.4 Post hoc analysis

To conduct pairwise t-test, independence, normality and equality of variance of each group must be satisfied. These assumptions were checked before ANOVA test, we assume these assumptions and continue the t-tests. Since there are 4 groups of smoking mother category, K = $4(4 - 1)/2 = 6$ times pairwise of t-test are performed. The significance level is set at 5% $\alpha = 0.05$. The null hypotesis of each test is that the means of two categories are equal $H_0 : \mu_i = \mu_j$ for some i and j where $i \neq j$. Table 5 represents the calculation results of the raw p-value, whereas table 6 represents p-value after being adjusted with Bonferroni correction.

	Never	Once did, not now	Smokes now
Once did, not now	0.359		
Smokes now	4.38×10^{-15}	6.10×10^{-8}	
Until current pregnancy	0.955	0.509	5.64×10^{6}

Table 4: Summary of the multiple two-sample t-tests without p-value adjustment

Through table 4, it can be seen the results of the multiple pairwise t-tests performed without p-value adjustment. The null hypothesis is rejected if the estimated p-value for each pairwise t-test is less than the fixed significance level, which is $\alpha = 0.05$ and from the result can be seen that for six pairs, a two sample t-test is performed. It can be observe that out of 6 test, 3 p-values are less that the α value. Hence, these 3 test are rejected the null hypothesis and it is claimed that they have a significant difference in the mean values of the resulting babies weight between the two smoking mother categories at the significance level of 0.05. For remaining 3 categories, the p-value is greater than α and the null hypothesis is not rejected for this pairwise t-test. The category that have p-value greater than significant level is between "Once did, not now" and "Never", between "Until current pregnancy" and "Never", lastly between "Until current pregnancy" and "Once did, not now".

	Never	Once did, not now	Smokes now
Once did, not now	1.00		
Smokes now	2.62×10^{-14}	3.66×10^{-7}	
Until current pregnancy	1.00	1.00	3.38×10^{-5}

Table 5: Summary of the multiple t-test with Bonferroni's p-value adjustment

In table 5, the Bonferroni correction adjusts the p-value by multiplying the raw p-values with the number of comparisons $p - value* = p - value \times 6$. After applied Bonferroni method, it can see that with the same 3 pairs of t-tests out of 6 pairs, its associated p-value is less than the fixed α value of 0.05. Hence, we reject the null hypothesis of these 3 tests, which claims that there are differences in the mean value of the resulting times between these categories.

	Never	Once did, not now	Smokes now
Once did, not now	0.796		
Smokes now	0	0	
Until current pregnancy	1.00	0.912	0

Table 6: Summary of the multiple t-test with Tukey's p-value adjustment

The result of p-value adjustment with Tukey method it shows in tabel 6. It shows the same result as Bonferroni, 3 pairs out of 6 pairs reject the null hypothesis. Table 7 shows the confidence interval for Tukey procedure.

Categories pair	Lwr	Upr
Once did, not now-Never	-3.15	6.63
Smokes now-Never	-11.5	-5.90
Until current pregnancy-Never	-4.82	5.04
Smokes now-once did, not now	-15.4	-5.52
Until current pregnancy-Once did, not now-Never	-8.02	4.75
Until current pregnancy-smokes now	3.84	13.8

Table 7: The confidence Interval of 6 pair variables

Table 7 gives the results of the Tukey confidence interval. There is none of zero value is within the interval. It means that all the variables are statistically significantly different from zero and also the null hypothesis is rejected.

5 Summary

In this project, the dataset is provided by the lecturers of Introductory Case Studies at TU Dortmund University in the summer semester of 2023. The dataset has observations of 1236 sample size data. The dataset has 23 independent variables, including the baby's weight and the smoking mother category, divided into five categories. Since the category "Unknown" is not provided any information about the smoking mother's history, then it can be deleted. The purpose of this project was to compare the baby's weight of four history smoking mother categories. ANOVA was conducted to check the average between the categories was equal. Another object was to pair two smoking mother categories and compare them to check they have equal means. To compare the difference between before and after correction when making multiple comparisons was one of the main objectives. Three assumptions which are independence, normality, and equality of variance, must be fulfilled to conduct ANOVA. The dataset in this project was collected randomly, so it is consider to satisfy independence. Q-Q plot was used to check normality, and since the data are generally distributed on the red line, it is supposed to satisfy normality. Boxplot and Levene's tests were used to check the equality of variance. The p-value from the Levene test was greater than the significant value that is satisfied with homogeneity. ANOVA tested the hypothesis that all groups had the same mean. As a result, the hypothesis was rejected and the post hoc analysis was performed. Multiple comparisons were conducted by pairing two groups. Bonferroni correction and Tukey procedure were used. The p- values before and after correction were compared and it can be seen that the p-value was the same. It was also the same result for Tukey methods. The variables were significantly different from zero and also the zero value for the variables is not in the interval. In the future, we can include variables such as the sex, age and mother's height to gain a better understanding of whether it affects their baby's weight.

Bibliography

Dahl, D. and Scott, D. (2019), *ggpubr: 'ggplot2' Based Publication Ready Plots.*

Davey, A. J. H. and Doncaster, C. P. (2007), *Analysis of variance and covariance: How to choose and construct models for the life sciences*, Cambridge University Press.

Dodge, Y. (2010), *The Concise Encyclopedia of Statistics*, Springer Science Business Media, 2008.

Graves, S. and PeterPiepho, H. (2019), *multcompView: Visualizations of Paired Comparisons.*

Grolemund, G. and Wickham, H. (2017), *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, O'Reilly Media, Canada.

Hastie, T. and Tibshirani, R. (2008), *The Elements of Statistical Learning*, Springer, California.

Jahans, C. (2019), *R companion to elementary applied statistics*, CRC, New York.

Kassambara, A. (2022), *ggpubr: 'ggplot2' Based Publication Ready Plots.*

Muth, J. (2014), *Basic Statistics and Pharmaceutical Statistical Applications*, CRC, New York.

R Development Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ross, S. M. (2010), *Introductory Statistics.*, Elsevier, The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, UK.

Roussas, G. G. (2003), *An introduction to probability and statistical inference*, Academic Press.

Selvamuthu, D. and Das, D. (2018), *Introduction to Statistical Methods, Desig of Experiments and Statistical Quality Control*, Springer Nature Singapore Pte Ltd.

Statlabs (2002), 'Maternal smoking and infant health i'. visited on 2023-06-01.

Verzani, J. (2014), *Using R for Introductory Statistics*, CRC.

Appendix

A Additional figures

B Additional tables