TU Dortmund

Introductory Case Studies

Project 2: Comparison of multiple distributions

Lecturers:
Prof. Dr. Katja Ickstadt
M. Sc. Zeyu Ding
M. Sc. Yassine Talleb

Author: Mohammad Sakhawat Hossain

Matriculation No: 231838

Group number: 10

Group members: Md Shahabub Alam, Shahed Iqbal Chowdhury, Hasan Zamil Ahmed, Sazeda Sultana

June 12, 2023

Contents

1 Introduction

Maternal smoking during pregnancy has long been recognized as a significant public health issue because of its potential adverse effects on fetal development and neonatal outcomes. Maternal smoking which is one crucial factor among various factors influencing the health of newborns has consistently emerged as an important determinant of birth weight with implications for both short and long term health outcomes. The association between maternal smoking and reduced birth weight has been extensively studied. However, it is still necessary to delve deeper into the nuanced relationship between smoking habits and neonatal weight outcomes, specially considering different smoking history (Knopik et al., 2016).

The goal of this project is to look at how different maternal smoking conditions are responsible to change the birth weight of neonates. Initially, descriptive statistics such as central tendency and dispersion are used to analyze the distribution of birth weight (wt) and smoking history ($smoke$) variables and performed a global test to assess differences in birth weight between the different smoking categories. To analyze the pairwise differences between the weights of different categories, we consider all pairs of categories and conduct a two-sample t-test for each pair. While performing multiple tests, such as t-test, there is an increased risk of false positive. To address this issue, we perform two correction methods such as Bonferroni and Tukey's HSD test. Finally, the findings were compared between the two correction methods and the non-adjusted tests, by providing a reasonable explanation for any observed differences.

In section 2, the data set and the project objectives are explained in detail. The data collection method, data quality, data preprocessing, data type as well as data size are described here. Section 3 discusses the statistical methods such as Q-Q plot, Levene test, analysis of variance (ANOVA), pairwise t-tests, Tukey's HSD test and Bonferroni method that are used for data analysis in this project. We use these statistical approaches in section 4, and the results of different statistical tests are thoroughly interpreted and analyzed. Section 5 presents the result and an in-depth summary of the project, followed by a discussion of potential future research on this data set.

2 Problem Statement

2.1 Description of the data set

The dataset utilized in this study is provided by the esteemed instructors of the "Introductory Case Studies" course at TU Dortmund, during the summer of 2023. Notably, the data has been meticulously collected from a reputable webpage affiliated with the distinguished Department of Statistics at the University of California, Berkeley (Berkeley Statistics, 2023).

The original dataset comprises 1236 observations and 23 independent variables. However, the provided dataset, named 'babies.csv' contains 1236 observations but only includes two variables, namely, *wt* and *smoke*. The variable *wt* is a continuous variable denoting the weight of newborn babies in ounces. On the other hand, the variable *smoke* is a categorical variable characterizing the smoking history of mothers under five distinct conditions: 0 = never, 1 = currently smoking, 2 = smoked until the current pregnancy, 3 = smoked previously but not presently, and 9 = unknown. Each of these conditions corresponds to 540, 481, 95, 100, and 10 observations, respectively. We removed 10 missing values and thus 1226 observations are used in our analysis.

2.2 Project objective

The main objective of this report is to apply statistical methods to examine and analyze the disparity in weight measured in ounces among new-born babies with a focus on five distinct smoking conditions of their respective mothers. At first, quantile-quantile plot (Q-Q plot) is used to identify the normality of the data and then Levene test is used to verify the homogeneity of variance. Since the normality assumptions and homogeneity of variance are observed, therefore one-way analysis of variance (ANOVA) is performed. We also use pairwise t-test to see whether there is a pairwise difference in the categories in terms of weight. To analyze the results of these tests, we employ the p-value approach. Finally, we use the Bonferroni method and Tukey's HSD test to deal with problems caused by multiple tests. We conclude our study by conducting a comparative analysis of the outcomes obtained from employing these tests and Bonferroni correction.

3 Statistical Methods

In this section, several statistics methods are introduced which are later used in our analysis. The software R (R Core Team, 2022), ggpubr (Kassambara, 2020), ggplot (Wickham, 2016) and tidyverse (Wickham et al., 2019) are used for data analysis and visualization. To do statistical analysis and summary statistics skimr (Waring et al., 2022), rstatix (Kassambara, 2021) and car (Fox and Weisberg, 2019) packages are utilized.

3.1 Null hypothesis(H_0) and alternative hypothesis (H_1)

Statistical hypothesis tests are the foundation of many statistical analyses and having a gainful insight into hypothesis testing is crucial. The data is analyzed with the assumption of particular outcome and then using statistical methods, we either reject or confirm the assumption. This assumption about the outcome is referred to as a hypothesis, and the statistical hypothesis tests are used to test these hypothesis. To accurately reflect the question that the tester wishes to answer, the hypothesis test must be carefully designed. There are two mutually exclusive hypothesis in a statistical hypothesis testing. First one is null hypothesis H_0 which aim to test and maintain if strong evidence isn't found against it. The alternative hypothesis, denoted as H_1 presents a statement that directly opposes the null hypothesis. If null hypothesis H_0 isn't true then the alternate hypothesis H_1 must be true and vice versa. If we don't reject a hypothesis it doesn't imply that the assumption made is correct rather it simply means that there is no evidence to prove the contrary (Banerjee et al., 2009).

3.2 Significance level (α) and p-value

The significance level is denoted by α and defined as - the null hypothesis is indeed true but there is a probability to reject it falsely. Researchers or statistician define the significance level before conducting the test. The significance level is 0.05 means there is a 5% chance of taking alternative hypothesis if null hypothesis is true. In our report, we will use 5%/0.05 significance level (Wasserman, 2010).

p-value method is one of the techniques to draw a conclusion whether null hypothesis can be rejected or not. When the assumption of null hypothesis is valid, the p-value corresponds to the probability of obtaining test results that are equal to or more extreme

than the observed result. Conclusion about a particular hypothesis depends on p-value and significance level. When the chosen significance level is greater than the obtained p-value ($\alpha > p$) then we reject null hypothesis, indicating a statistically significant result. If the significance level is less than the p-value ($\alpha < p$) then we fail to reject null hypothesis and hence the outcome is considered statistically insignificant (Du Prel et al., 2009).

3.3 Statistical tests and assumptions

In this report two different tests, namely pairwise t-test and one-way ANOVA are used as inferential statistics. To adjust the result found in ANOVA and t-test we use Bonferroni correction and Tukey's HSD test. A comprehensive discussion regarding the details of these tests will be provided in subsequent sections. To proceed with the inference about the means of population certain assumptions have to be held for these tests (Black et al., 2018, p. 409).

- Populations should be normally distributed.
- Populations should have equal variances.
- Observations in the populations should be independent with each other.

3.4 Q-Q Plot

The quantile-quantile plot (Q-Q plot) is a graphical method for determining whether the distributional properties of the sample come from a theoretical distribution like the normal distribution. Considering sample data $y_1, y_2, ..., y_n$ are sorted in ascending order $y_{(1)}, y_{(2)}, ..., y_{(n)}$, they are referred to as observed quantiles. The probability points p_i, for $i = 1, 2, 3, ., n$ is calculated using the given formula:

$$pi = \begin{cases} \frac{i-\frac{3}{8}}{n+\frac{1}{4}} & \text{if } n \leq 10 \\ \frac{i-\frac{1}{2}}{n} & \text{if } n > 10. \end{cases}$$

The theoretical quantiles t_i, for their corresponding sorted sample quantile y_i are computed using probability points pi. For $i = 1, 2, .., n$ we then find t_i , the theoretical quantile such that $P(T \leq t_i) = pi$ and $T \sim N(0, 1)$ i.e T is normally distributed with mean value 0 and variance 1. For plotting the graph, ordered pair (t_i, y_i) are used. On

the same graph a reference line is superimposed. In normal probability Q-Q plot, the sample mean is the y-intercept and the standard deviation is the slope of reference line. For non standardized data reference, line is of the form $y = \mu + z$ where z is the transformation computed using $(y - \bar{y})/s$, μ represents mean and s represents standard deviation of proposed theoretical normal distribution. For standardized data, the reference line takes the form $y = t$. Once the plot is ready and the reference line is superimposed on to the plot, it is possible to visually inspect if the data points follow the theoretical distribution. If the points follow the reference line, then it is highly likely that they are normally distributed. Deviations from the reference line imply a deviation from normality and have a noticeable concave trend. The presence of such concave trends leads to skewness in the plot (Hay-Jahans, 2019, p. 147-152).

3.5 Levene test for variance equality

The Levene test is an inferential test which is used to assess whether the variances of two or more groups or samples are equal or not. This test is important as it is an assumption in many statistical tests, such as t-test and ANOVA. The null hypothesis denoted by H_0 states that the population variances of all k groups are equal and alternative hypothesis denoted by H_1 asserts that at least one group's population variance is not equal to the others.The mathematical representation of the hypothesis is given below:

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$$
$$H_1 : \sigma_i^2 \neq \sigma_j^2 \text{ (for at least one pair } i \text{ and } j)$$

Let j ranges from 1 to k $(j = 1, 2, ..., k)$, representing each individual sample. n_j denotes the size of the j^{th} sample, and y_{ij} represents the i^{th} observation within the j^{th} sample. The mean of the j^{th} sample is denoted as $\bar{y_j}$. Additionally, the absolute deviations is calculated as $\Delta_{ij} = |y_{ij} - \bar{y_j}|$

Consider $n = \sum_{j=1}^{k}(n_j)$, which represents the total size of all samples combined. Let, $\bar{\Delta}$ represent the average value of the absolute deviations, the sample mean and variance is denoted by $\bar{\Delta_j}$ and $s^2_{\Delta_j}$ respectively, of the absolute deviations. Then the test statistic is as follows:

$$F^* = \frac{\sum_{j=1}^{k} n_j(\bar{\Delta_j} - \bar{\Delta})^2(n-k)}{\sum_{j=1}^{k}(n_j - 1)s^2_{\Delta_j}(k-1)}$$

Here, p-value $= P(F \geq F^*)$ and $F^* \sim F(k-1, n-k)$. If p-value is greater than the assumed significance level 0.05 then we don not reject the null hypothesis i.e. equal variances exist among the groups and vice versa (Hay-Jahans, 2019, p. 247-248).

3.6 Analysis of variance (ANOVA)

ANOVA is a statistical test which is used to determine the mean difference between more than two data samples or groups. ANOVA can be one way or two way test. In this report we will use one way ANOVA as we have only one quantitative independent variable. For two independent variable we use two-way ANOVA test. In ANOVA if we have k samples then the following hypothesis are tested:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$
$$H_1 : \text{At least one of the means is different from the others.}$$

When one of the sample mean is different than others, then we reject the null hypothesis and we can say there is significant difference among the groups or data samples. To conduct the test first we have to measure between and within group variances. Variances between the groups (SSB) can be calculated by subtracting variances within groups (SSW) from total variances (SST)

$$SST = \sum_{i=1}^{n}\sum_{j=1}^{k}(x_{i,j} - \bar{X})^2 \text{ and } SSW = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{i,j} - \bar{x_j})^2.$$

Here, $\bar{X}$ is the grand mean which can be defined as:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} x_{i,j}.$$

x_{ij} is the individual data points of all groups, k is the number of groups or data samples and $\bar{x_j}$ is the mean value of group j. n is the total number of sample size. Next, mean square between group (MSB) and mean square within group (MSW) is calculated by using the following formula:

$$MSW = \frac{SSW}{\sum_{j=1}^{k}(n_j - 1)} = \frac{SSW}{n-k} = \frac{SSW}{df_w} \text{ and } MSB = \frac{SSB}{k-1} = \frac{SSB}{df_b}$$

Here, degree of freedom is referred to as the maximum number of logically independent values that have the freedom to vary. The number of degrees of freedom between groups, denoted as df_b, is equal to $k-1$, where k represents the number of groups. On the other hand, df_w is the degrees of freedom within groups that correspond to $n-k$ where n represents the total sample size (Tukey, 1949, p. 232-242). Finally F-statistics F^* is the ratio of the two variances:

$$F^* = \frac{MSB}{MSW}$$

To check whether the null hypothesis should be rejected or not, p-value method can be used. In ANOVA p-value corresponds to the probability $P(F \geq F^*|H_0)$ where F is random variable comes from $F_{(df_w, df_b)}$ distribution in significance level α. When the chosen significance level is greater than the obtained p-value ($\alpha > p$) then we reject null hypothesis, indicating a statistically significant result. If the significance level is less than the p-value ($\alpha < p$) then we fail to reject null hypothesis and hence the outcome is considered statistically insignificant (Black et al., 2018, p. 406-411).

3.7 Pairwise t-test

By comparing the mean values of the groups in the preceding part, the ANOVA test merely indicated whether it reject the null hypothesis or not; no information is provided as to which two groups had similar mean values and which did not. To find out whether there is a significant difference in the mean values between pairwise groups or categories pairwise t-test should be done. Similar to the ANOVA test, this test's assumptions are already provided at the beginning of this section. For all pairs i, j where $i, j = 1, ..., k$ the following is the hypothesis for this test:

$$H_0 : \mu_i = \mu_j$$
$$H_1 : \mu_i \neq \mu_j \text{ with } i \neq j$$

In order to obtain t-statistic, we need to compute a pooled sample standard deviation (S_p) which is a weighted average of standard deviation taken from two or more than two independent groups or data samples. For more than two independent data samples the

formula of pooled SD can be defined as follows:

$$S_p^2 = \frac{(n_1 - 1)\ s_1^2 + (n_2 - 1)\ s_2^2 + .. + (n_k - 1)\ s_k^2}{n_1 + n_2 + .. + n_k - k}$$

If $n_1, n_2.....n_k$ are the sample sizes of group k and $s_1, s_2, ...s_k$ are the standard deviation of k groups then assuming the pooled SD same ($\sigma_1 = \sigma_2$) we can write the pairwise-test formula as:

$$t^* = \frac{\bar{x_i} - \bar{x_j}}{\sqrt{\frac{(n_1-1)\ s_1^2+(n_2-1)\ s_2^2+..+(n_k-1)\ s_k^2}{n_1+n_2+..+n_k-k}}\sqrt{(\frac{1}{n_i} + \frac{1}{n_j})}}$$

where $\bar{x_i} - \bar{x_j}$ is the mean difference between two groups.

After having the t-statistic, in order to check whether we should reject null hypothesis or not, p-value approach can be used. In two-tailed pairwise t-test p-value corresponds to the probability $2P(T \geq | t^* | | H_0)$ where T is a random variable comes from $T_{n_1+n_2+...n_k-k}$ distribution in significance level α. If the chosen significance level is greater than the obtained p-value ($\alpha > p$) then we reject null hypothesis, indicating a statistically significant result. If the significance level is less than the p-value ($\alpha < p$) then we do not reject null hypothesis and hence the outcome is considered statistically insignificant (Hay-Jahans, 2019, p. 261).

3.8 Bonferroni method

Whenever we reject the null hypothesis wrongly then this type of instance is called type I error or false positive. On the other hand, whenever we don't reject the null hypothesis wrongly then this type of instance is called type II error or false negative (El-gohary, 2019). When several dependent or independent statistical test are performed at a time the false positive rate increases substantially and therefore the Bonferroni is used with a view to reduce the instance of false positive rate. Let $H_1, ..., H_m$ is a family of hypothesis and there corresponding p-values $p_1, ..., p_m$. Let m_0 is the number of true null hypothesis where m is the total number of null hypothesis then family wise error rate (FWER) is the probability of making at least one type I error i.e probability of rejecting at least one true H_i. We can calculate the family wise error rate as follows:

$$FWER = 1 - (1 - \alpha)^m$$

where m is number of tests and α is significance level.

To avoid the type I error, we can test each hypothesis by using alternate significance level i.e, α/m. Therefore, we can use the adjusted α and compare it with the p-value. Adjusted α value can be calculated as follows:

$$\alpha_{adjusted} = \alpha/m$$

Although this method significantly reduces type I error but increases chance to the vulnerability of type II error (Field, 2013, p. 153-154)

3.9 Tukey's HSD test and confidence interval

Tukey's HSD (Honestly Significant Difference) is a post-hoc test that can be used for performing multiple pairwise comparisons between group means by controlling the overall type 1 error rate. For all pairs i, j where $i, j = 1, ..., k$ the hypothesis for this test can be stated as follows:

$$H_0 : \mu_i = \mu_j$$
$$H_1 : \mu_i \neq \mu_j \text{ with } i \neq j$$

To calculate Tukey's HSD the formula for the test statistic can be defined as:

$$Q_{i,j} = \frac{\sqrt{2}(\mu_i - \mu_j)}{s\sqrt{\frac{n_i+n_j}{n_i n_j}}} \sim q(k, n-k)$$

Here, μ_i and μ_j are the means of two different group i and j respectively, s is standard error, n_i and n_j indicates sample size of two distinct groups.

To calculate p-value $= P(q \geq Q_{i,j})$, this test statistic can be used where q is the studentized range distribution. If p value is less than the significance level 0.05, we can conclude that a significance difference can be observerd between two particular groups and vice versa.

A confidence interval (CI) is an important statistical tool that provides a range of values within which the actual population parameter can be reasonably predicted to lie. 95%

confidence interval is commonly used which indicates that we are 95% confident that the true population mean weight lies between 120 ounces and 130 ounces if the mean weight of a population can be expressed between 120 and 130 ounces.

If standard deviation is not available we can use alternative methods to calculate CIs, such as Tukey's Honestly Significant Difference (HSD) test. The HSD test calculates confidence intervals for all pairwise comparisons of group means and focuses on the studentized range distribution or 'q' distribution.

If $\mu_i > \mu_j$ then Tukey's confidence interval is defined as follows:

$$\mu_i - \mu_j \pm \frac{1}{\sqrt{2}} c_\alpha \; s \sqrt{\frac{n_i + n_j}{n_i n_j}}$$

c_α is the critical value that represents the cutoff point in the distribution beyond which values are considered extreme. Here, $c_\alpha > 0$ such that

$$P(q \geq c_\alpha) = \alpha$$

with $q \sim q(k, n-k)$ indicates that q follows the q-distribution depends on k groups and $n-k$ degree of freedom where α represents a significance level. If the confidence interval includes the value associated with the null hypothesis (e.g., 0 for the difference in means), we fail to reject the null hypothesis. On the other hand, if the confidence interval does not include this value, it suggests that we have evidence to reject the null hypothesis (Hay-Jahans, 2019, p. 276).

4 Statistical Analysis

4.1 Descriptive analysis and checking assumptions

In this subsection, we check the assumptions and describe the data as well as the distribution for each category. Summary statistics for all 5 categories are represented in the table 2 on page 18 in the appendix section. According to this table and histogram in Figure 1, we observe that the babies whose mother's smoking history is unknown (category 9) have a small range (68) but high variance among all categories where all the other variances are roughly same. Mean (126.70) and median (128) value is also

higher for this category. For Category 1, most of the babies weight lies between 90 and 140 ounces with the smallest mean value 114.11 ounces.

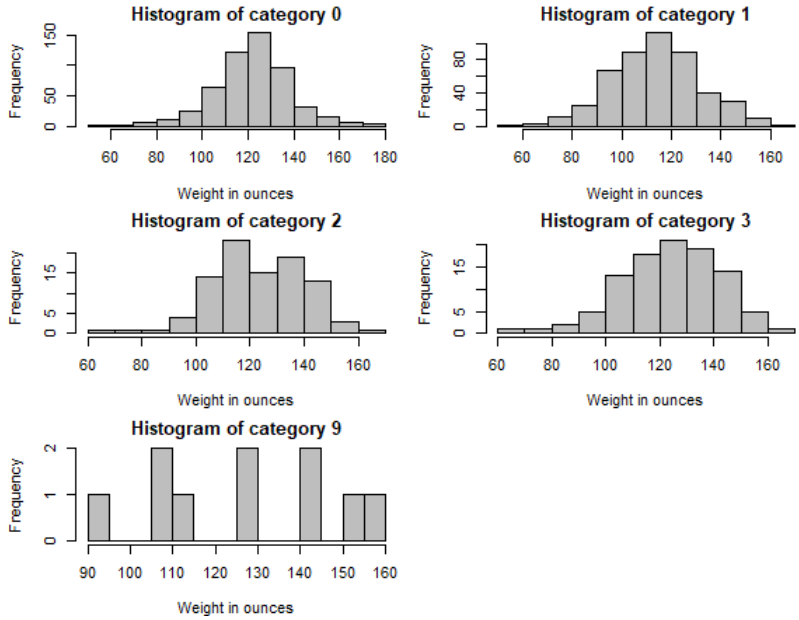

Figure 1: Frequency Distribution of Weight in Ounces.

The average values of 122.86, 123.08, and 124.63 for categories 0, 2, and 3 respectively show that these three categories exhibit similar characteristics. This similarity is also observed in the median values. The histogram analysis reveals that a significant proportion of babies in categories 1 and 2 have weights falling within the range of 100 to 150.

In order to visualize whether the samples are normally distributed or not, Q-Q plots are used. In Figure 2, One individual Q-Q plot is shown for each category. Category 1 seems to follow a normal distribution since its quantile is approximately on the reference line. For categories 2, 3, and 9, the majority of data points align closely with the reference line, with only a few exceptions. In contrast, few data points in category 0 are far from the reference line, indicating some babies have exceptionally low and high weights within this category. The substantial presence of data points closely touching the reference line in majority of categories leads to the assumption that all data points show a distribution that can be approximated as normal.

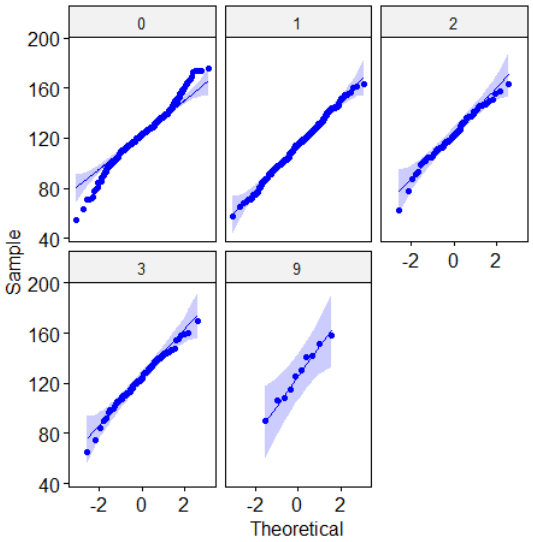

Figure 2: QQ Plots of weight by smoke category.

To verify the homogeneity assumptions, we use Levene test. From the coding part, we see all the p values (0.1127) are larger than the assumed significance level 0.05. Based on the aforementioned findings, we don't reject the null hypothesis and can conclude that weights of all the babies exhibit homogeneous variance.

Furthermore, the weight of neonates in the given data set are unrelated to one another, so the occurrence of one property has no effect on another. It implies that we can assume that the independence assumptions are likewise true.

4.2 Global test

In this section, we will use a global test called ANOVA to investigate whether the weight differs between the five categories or not. In the method sections, the underlying assumptions for this test are stated and checked in the preceding section. The hypothesis for this test can be assumed as follows:

- Null Hypothesis, H_0 : For all five categories, the average weight (in ounces) is the same.
- Alternate Hypothesis, H_1: The average weight taken (in ounces) differs in five categories.

We can see degrees of freedom for the numerator (DF_n), degrees of freedom for the denominator (DF_d), F-statistic, and p-value from the Table 1. By using the p-value

method, we can decide whether the null hypothesis will be rejected or not. Since the p-value is less than the presumed significance level (0.05), we reject the null hypothesis. Rejecting the null hypothesis means there is a statistically significant difference among the five categories in terms of weight. However, to find out how each pair of categories differ and specifically which pair of categories cause the rejection, we will use a pairwise t-test in the next section.

Variable	DF_n	DF_d	F Value	p-value
Smoke	4	1221	19.62	1.15×10^{-15}

Table 1: Output of the ANOVA test.

4.3 Pairwise t-test, adjusting and comparing the results

Pairwise t-test are performed for ten pairs of categories to determine whether there are pairwise weight differences between the two categories. Since it is assumed that all groups come from populations with a common standard deviation, by using pool.sd = TRUE statement in pairwise t-test function a common variance estimator for all groups is obtained. The underlying assumptions are already checked, and the hypothesis can be defined as follows:

- Null Hypothesis, H_0 : The mean value of weight does not differ between any two categories.
- Alternate Hypothesis, H_1: Between the two categories, there is a difference in the mean value of weight.

The results of 10 pairwise t-tests are demonstrated in Table 3 on page 18 in the appendix section. We observe that 1-0, 2-1, 3-1 and 9-1 pairs of categories have smaller p-value than the significance level (0.05). Therefore, we can reject the null hypothesis for these 4 pairs and thus conclude that there are significant differences between these categories in terms of average weight. In contrast, all other 6 pairs of categories have greater p-value than the significance level, so we cannot reject the null hypothesis, thus there is no difference in average weight between these categories.

Since we conduct the 10 statistical analyses on the same sample data, the family-wise error rate (FWER) increases, and FWER can be calculated as follows:

$$FWER = 1 - (1 - \alpha)^m = 1 - (1 - 0.05)^{10} = 0.40$$

So, the chance of erroneously rejecting the null hypothesis at least once among the family of analyses is equal to 40%. As mentioned earlier, to control the inflated family-wise error rate in multiple testing, we should make an adjustment. In this report, Bonferroni and Tukey's HSD are used to control the FWER. From Table 4 on page 18 in the appendix section, we see all the adjusted p-values with the Bonferroni method. Here, the p-value for 9-1 category (0.26) is larger than previous p-value in t-test (.026). Three pairs of categories 1-0, 2-1 and 3-1 have smaller p-values than significance level 0.05 while all other pairs have larger p-value than the significance level.

To have a closer look on specific pair of categories we also conduct Tukey's HSD test. From Table 4 on page 19 in the appendix section, we observe that three pairs of categories such as 1-0, 2-1 and 3-1 have p-value smaller than 0.05. If we compare with the previous t-test we see one pair 9-1 have larger p-value while conducting Tukey's test. Consequently, we can deduce that there are statistically significant differences in the average weight of neonates among three pairs of categories: 1-0, 2-1, and 3-1.

At a confidence level of 0.95, it is observed that the lower confidence level (LCL) and upper confidence level (UPL) for 0-1 category are -11.78 and -5.73 which doesn't include zero. Similarly, we also observe the same LCL and UPL for the pair of categories 2-1 and 3-1 are (3.56, 14.39) and (5.22, 15.82) respectively which doesn't include any zero and thus we reject null hypothesis and can conclude that there are significant differences among these pair of categories while no significant differences for the rest of pair of categories.

By comparing the t-test with Bonferroni correction and Tukey's HSD test, a noticeable pattern can be observed. Only one pair of groups, namely 9-1, exhibits larger p-values in both the Bonferroni correction and Tukey's HSD test. In contrast, the results for the three pairs of categories, 1-0, 2-1, and 3-1, obtained from the t-test, Bonferroni correction, and Tukey's HSD test, reveal that the p-values for these three categories are lower than the assumed significance level of 0.05. Conversely, all other pairs demonstrate larger p-values exceeding 0.05. Consequently, we can conclude that significant differences exist in the weights of babies between mothers who never smoked and currently smoke (1-0), mothers who smoked until the current pregnancy and currently smoke (2-1), and mothers who smoked in the past but not currently and currently smoke (3-1).

5 Summary

In this project, a comparison of multiple tests was conducted to investigate the relationship between maternal smoking status and its impact on infant weight. The data set was collected from the webpage affiliated with the statistics department at the University of California, Berkeley, and put together by the instructors of the course Introductory Case Studies at TU Dortmund University during the summer, 2023. The given data set includes 1236 observations with a categorical variable *smoke* and continuous variable *wt* which denotes the weight of the neonates. The categorical variable comprised five distinct categories. Category 0 referred to mothers who never smoked. Category 1 represented mothers who reported currently smoking. Category 2 included mothers who reported smoking until their current pregnancy. Category 3 indicated mothers who reported having smoked at some point in the past but had since ceased smoking. Lastly, category 9 encompassed mothers for whom the history of their smoking habit was unknown.

Initially, we used an ANOVA test to see any difference in categories that were varying in terms of average weight. As the p-value is less than the assumed significance level, therefore we concluded that a difference between categories existed. To investigate the pairwise difference between the categories we used a pairwise t-test, which gives us smaller p-value than assumed significance level 0.05 in 4 pairs of categories such as 1-0,2-1, 3-1 and 9-1. Therefore, we concluded that there were significant differences between these pairs of categories. Since this task requires performing multiple statistical tests, the risk of a family-wise error rate increases. To address this issue, we introduced two correction methods such as Bonferroni and Tukey's HSD, where both methods give roughly the same result as pariwise t-test except one pair of category 9-1 had larger p-value for t-test.

Our analysis focused solely on the variables smoke and birth weight. There may be other confounding factors, such as maternal health, socioeconomic status, or prenatal care, which were not taken into account. In conclusion, this study provides evidence of a significant relationship between maternal smoking and babie's birth weight, with infants born to smoking mothers displaying lower birth weights. However, we need to consider the limitations and potential confounding factors as well. Further research is needed to better understand the complex interactions between maternal smoking, birth weight, and other relevant factors.

Bibliography

Amitav Banerjee, UB Chitnis, SL Jadhav, JS Bhawalkar, and S Chaudhury. Hypothesis testing, type i and type ii errors. *Industrial psychiatry journal*, 18(2):127, 2009.

Berkeley Statistics. STAT LABS: Data. `https://www.stat.berkeley.edu/users/statlabs/labs.html`, 2023. [Visited on 03-06-2023].

Ken Black, John Asafu-Adjaye, Paul F Burke, Nazim Khan, Gerard King, Nelson Perera, Andrew Papadimos, Carl Sherwood, and Saleh Ahmed Wasimi. *Business analytics and statistics*. John Wiley & Sons Australia, Limited, 2018.

Jean-Baptist Du Prel, Gerhard Hommel, Bernd Röhrig, and Maria Blettner. Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, 106(19):335, 2009.

Tarek M El-gohary. Hypothesis testing, type i and type ii errors: Expert discussion with didactic clinical scenarios. *International Journal of Health and Rehabilitation Sciences (IJHRS)*, 8(3):132, 2019.

Andy Field. *Discovering statistics using IBM SPSS statistics*. sage, 2013.

John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition, 2019. URL `https://socialsciences.mcmaster.ca/jfox/Books/Companion/`.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. CRC Press,New York, 2019. ISBN 9780429448294. doi: 10.1201/9780429448294.

Alboukadel Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*, 2020. URL `https://CRAN.R-project.org/package=ggpubr`. R package version 0.4.0.

Alboukadel Kassambara. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*, 2021. URL `https://CRAN.R-project.org/package=rstatix`. R package version 0.7.0.

Valerie S Knopik, Kristine Marceau, Rohan HC Palmer, Taylor F Smith, and Andrew C Heath. Maternal smoking during pregnancy and offspring birth weight: a genetically-informed approach comparing multiple raters. *Behavior genetics*, 46:353–364, 2016.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

John W Tukey. One degree of freedom for non-additivity. *Biometrics*, 5(3), 1949.

Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu, and Shannon Ellis. *skimr: Compact and Flexible Summaries of Data*, 2022. URL https://CRAN.R-project.org/package=skimr. R package version 2.1.5.

Larry Wasserman. *All of statistics : a concise course in statistical inference*. Springer, New York, 2010. ISBN 9781441923226 1441923225. URL http://www.amazon.de/All-Statistics-Statistical-Inference-Springer/dp/1441923225/ref=sr$_{12}$?$ie = UTF8qid = 1356099149sr = 8 - 2$.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Appendix

A Additional tables

Category	variable	no. of neonates	min	max	mean	median	IQR	SD
0	weight	540.00	55.00	176.00	122.86	124.00	18.50	17.06
1	weight	481.00	58.00	163.00	114.11	115.00	24.00	17.97
2	weight	95.00	62.00	163.00	123.08	122.00	24.50	17.80
3	weight	100.00	65.00	170.00	124.63	124.50	26.00	18.57
9	weight	10.00	90.00	158.00	126.70	128.00	32.00	21.81

Table 2: Summary statistic of different categories (weight measured in ounces).

	0	1	2	3
1	5.6×10^{-15}	-	-	-
2	0.91	6.4×10^{-6}	-	-
3	0.36	7×10^{-8}	0.54	-
9	0.50	0.026	0.538	0.724

Table 3: p-value of pairwise comparison without adjustment method.

	0	1	2	3
1	5.6×10^{-14}	-	-	-
2	1.00	6.4×10^{-5}	-	-
3	1.00	7×10^{-7}	1.00	-
9	1.00	0.26	1.00	1.00

Table 4: p-value of pairwise comparison with Bonferroni adjustment method.

Pairs of categories	diff	lwr	upr	p adj
1-0	-8.75	-11.78	-5.73	0.0000000
2-0	0.22	-5.14	5.59	0.9999624
3-0	1.77	-3.48	7.02	0.8889099
9-0	3.84	-11.54	19.22	0.9604221
2-1	8.98	3.56	14.39	0.0000631
3-1	10.52	5.22	15.82	0.0000007
9-1	12.59	-2.81	27.99	0.1680084
3-2	1.55	-5.36	8.45	0.9733043
9-2	3.62	-12.41	19.64	0.9725115
9-3	2.07	-13.92	18.06	0.9966488

Table 5: p-value of Tukey's adjustment method and confidence interval.