TU Dortmund

Introductory Case Studies

# Project 2: Comparison of multiple distributions

July 27, 2023

# Contents

# 1 Introduction

This statistical report intends to investigate the relationship between maternal smoking and infant weight, focusing on whether different smoking conditions lead to changes in the weight of neonates in distinct groups. The "Babies" dataset, consisting of 1236 samples and 23 independent variables of newborns and individual mothers, contains the necessary data for our analysis. The dataset contains variables such as birth weight, infant survival, date of birth, gender, mother ethnicity, age, education level, height, weight, and smoking status (Berkeley Statistics, 2023).

Our problem-solving approach employs various statistical techniques to address our research questions. Initially, descriptive statistics will be utilized to summarize the distribution of the variables of interest. We will consider the data count for discrete variables, whereas for continuous variables, we will investigate the distribution and central tendency of the data across various groups.

To determine whether the birth weights of babies vary by category, we will conduct a global test. In addition, we will investigate pairwise differences in birth weights by conducting two-sample $t$-tests for all possible category pairs. We will adjust the test results for multiple comparisons using the Bonferroni correction and Tukey's Honest Significant Difference (HSD) method. In addition, we will calculate Tukey's confidence interval to measure the precision of the detected differences.

The report will compare the results obtained from these correction methods to the unadjusted test and provide a reasonable explanation for the differences observed. The primary outcomes of our study will provide insight into the impact of maternal smoking on neonatal weight and ascertain whether particular smoking circumstances are linked to notable weight variations among distinct cohorts of infants.

Section 2 provides a detailed description of the dataset and its variables. The third section describes the methodology and statistical techniques used in our analysis, including hypothesis testing, Q-Q plots, one-way ANOVA, two-sample $t$-tests , multiple comparisons, Bonferroni correction, Tukey's procedure, confidence intervals, and Levene's test. Section 4 presents the key findings, including descriptive statistics and global and pairwise test results. We also discuss the implications of the findings. Section 5 concludes with concluding remarks and recommendations for future research.

# 2 Problem statement

The dataset titled "Babies" has been collected to investigate the correlation between maternal smoking and the weight of newborns. The dataset comprises data associated with individual mothers and their respective infants. The dataset consists of 1236 observations and 23 predictor variables, encompassing birth weight, birth date, gender, maternal ethnicity, age, educational attainment, height, weight, and smoking habits. For this research, we are considering only weight and smoking categories (Berkeley Statistics, 2023).

## 2.1 Dataset description

The dataset contains 1236 observations in total. For this research, we consider two variables (wt and smoke). It is important to note that the dataset contains 10 missing values that have been omitted from the analysis. After omitting 10 missing values the current dataset has 1226 observations. The general quality of the data is considered satisfactory. The value 999 indicates unknown data for the first variable, wt, which represents the birth weight of babies in ounces. This variable is of continuous numeric type, enabling quantitative weight distribution analysis. The second variable, smoke, represents the smoking history of mothers and is coded as follows: 0 for never smoking, 1 for currently smoking, 2 for until current pregnancy, 3 for having smoked in the past but not currently, and 9 for an unknown smoking history. This categorical variable permits the examination of the relationship between smoking status and variations in birth weight among neonates (Berkeley Statistics, 2023).

## 2.2 Project objective

This project aims to analyze the effect of maternal smoking on infant weight and determine whether different smoking conditions lead to variations in infant weight. Beginning with descriptive statistics, the project will examine the distribution of birth weight and smoking status variables. The data count will be assessed for discrete variables, whereas the distribution and central tendency of continuous variables will be analyzed across various smoking status groups. A global test will be conducted to determine if there are significant differences in the birth weights of babies between categories of smoking status. Two-sample tests will be used to conduct pairwise comparisons of birth weights,

and the results will be adjusted using the Bonferroni correction and Tukey's Honest Significant Difference (HSD) method. In addition, the confidence interval of Tukey will be calculated. The results of the adjusted test will be compared to those of the unadjusted test, and any differences observed will be explained. The necessary test assumptions will be verified, and a significance level of $\alpha = 0.05$ will be utilized. For each test, the null and alternative hypotheses will be clearly stated.

# 3 Statistical methods

In this section, we utilize the R software (R Core Team, 2022) and the following additional packages: skimr (Waring et al., 2022), tidyverse (Wickham et al., 2019), rstatix (Kassambara, 2021), ggplot2 (Wickham, 2016), ggpubr (Kassambara, 2020), gridExtra (Auguie, 2017), and car (Fox and Weisberg, 2019). These packages provide various statistical methods for analyzing and visualizing the data.

## 3.1 Hypothesis testing

Hypothesis testing is a method where a specific value for a parameter of the population probability distribution is selected. These values for a parameter are the mean, variance, or proportion.

### 3.1.1 Null hypothesis and Alternative hypothesis

The null hypothesis supports a specific assumption being tested. First, the value is selected as an initial null hypothesis and is denoted by $H_0$. This null hypothesis holds unless there is substantial evidence against it. So, if the null hypothesis is rejected, the alternative hypothesis denoted by $H_1$ is accepted. In addition, if the null hypothesis is failed to reject, it is not necessarily inferred that the null hypothesis is true. After defining the hypotheses, an appropriate test statistic is chosen based on the hypotheses (Newbold et al., 2013, p. 347).

### 3.1.2 Type-I and Type-II error

There are two types of errors present in statistical tests. The first one is a Type-I error. When the null hypothesis is true and the null hypothesis is rejected, it is called a Type-I

error. Moreover, Type-II is defined for the second one when the null hypothesis is not true, but the null hypothesis is not rejected (Heumann et al., 2016, p. 213).

### 3.1.3 Significance level

The significance level defined as $\alpha$ is the probability of rejecting the null hypothesis when it is true, also known as Type-I error. When the null hypothesis is rejected, the alternative hypothesis is accepted. The significance level is predetermined and usually set to 1 percent, 5 percent or more. The sample size and the unknown actual parameter values determine the probability of Type-II error. These values are present in the population being investigated (Heumann et al., 2016, p. 213).

### 3.1.4 $p$-value

The $p$-value represents the probability of obtaining results under the lowest alpha level at which the null hypothesis is rejected. A smaller $p$-value indicates more robust evidence against the null hypothesis. When the $p$-value is very small, it provides more substantial evidence against the null hypothesis. If the calculated $p$-value is less than or equal to the pre-specified significance level, the null hypothesis is rejected, and the alternative hypothesis is accepted. Contrarily, the null hypothesis is not rejected if the $p$-value exceeds $\alpha$ (Heumann et al., 2016, p. 215).

## 3.2 Q-Q plots

Quantile-quantile plots (abbreviated Q-Q plots) are a graphical tool used to assess the similarity between the distribution of an ordered sample and a theoretical distribution. Here, the quantiles of the ordered sample are compared to the quantiles of the theoretical distribution. The ordered samples are sorted in ascending order and the corresponding quantiles are then calculated. These quantiles are plotted against the expected quantiles from the theoretical distribution. The observed quantiles are first computed by sorting the $n$ sample from least to largest using $y_{(1)}$, $y_{(2)}$, $y_{(3)}$,..., $y_{(n)}$. Thus, plotting points $p_i$ are calculated as follows,

$$p_i = \begin{cases} (i - 3/8)/(n + 1/4) & \text{if } n \leq 10 \\ (i - 1/2)/n & \text{if } n > 10 \end{cases}$$

In this case, $n$ represents the sample size of the observed quantiles and $p_i$ represents the plotting point for the $i$-th member of the ordered sample. Based on theoretical quantiles $x_i$, ascending sample quantiles are calculated using the computed plotting points $p_i$. In this study, we consider theoretical quantiles $x_i$ such that $\mathrm{P}(X \leq x_i) = p_i$, where $X$ is a random variable and $X \sim N(0, 1)$. Ascending ordered sample quantiles are displayed on the $y$-axis, while theoretical quantiles are displayed on the $x$-axis. Here, a reference line is plotted using the first and third quantiles of the theoretical and ordered sample quantiles. Here, $a_1$ and $b_1$ are the first quantiles for the theoretical and ordered sample quantiles. In addition, $a_2$ and $b_2$ as the third quantiles for the theoretical and ordered sample quantiles, respectively. Now, it is possible to draw the reference line so that it goes through the $(a_1, b_1)$ and $(a_2, b_2)$ points (Hay-Jahans, 2019, p. 146-148).

## 3.3 One-way ANOVA

One-way analysis of variance (ANOVA) determines whether two or more independent groups determine whether there is statistical evidence that the relevant population means are significantly different. There are $k$ groups being analyzed, and $\mu_1, \mu_2, \mu_3, .., \mu_k$ represent the population means of these groups. The same notation will be used for the formalization of our hypothesis, which will be tested in the subsequent section as follows,

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k \qquad vs. \qquad H_1 : \mu_i \neq \mu_j \ ,$$

At least one pair $(i, j)$ exists where $i$ is not equal to $j$. Here, $k$ is denoted as the overall mean of all the observed responses by $\bar{y_j}$ and the $j^{th}$ sample variance is denoted by $s_j^2$.

The independence of the sample and population within each group is considered an underlying assumption. Additionally, it is assumed that the observed responses, denoted as $y_{ij}$, indicate independence within each respective sample. Furthermore, it is assumed that the underlying random variables of the populations correspond to a normal distribution. Finally, it is commonly assumed that the variances of samples and populations that belong to distinct groups are either equivalent or illustrate homogeneity. The test statistic used to assess the equality of the $k$ means derived from the mentioned assumptions is as follows,

$$F^* = \frac{\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2/(k-1)}{\sum_{j=1}^{k}(n_j - 1)s_j^2/(n-k)} \sim F(k-1, n-k)$$

The variables $k$ and $n$ denote the entire count of groups and the total number of observations that includes all groups. These symbols, namely $\bar{y}_j$, $\bar{y}$, and $s_j^2$, represent the average value of all measurements in the $j$-th category, the average value of all measurements taken collectively and the variance of the sample measurements in the $j$-th category respectively. It is imperative to highlight that these values are regarded as realized observations.

The observed value of the test statistic is represented by the symbol $F^*$. Therefore, it does not possess a standard $F$-distribution. Under the null hypothesis ($H_0$), the distribution of the test statistic in the population corresponds to an $F$-distribution. The test statistic, denoted by $F^*$, fits an $F$-distribution with $k-1$ degrees of freedom in the numerator and $n-k$ degrees of freedom in the denominator.

Therefore, the inequality $F^* > F(k-1, n-k)$ is employed to evaluate the statistical significance of the results by comparing the observed test statistic to the critical value from the $F$-distribution with the appropriate degrees of freedom.

In statistical hypothesis testing, if the computed $p$-value ($P(F \geq F^*)$) is greater than the predetermined significance level, usually set at 0.05, it implies insufficient evidence to reject the null hypothesis. This implies that the data that has been observed is in agreement with the null hypothesis, and any observed variations or impacts can be determined by probability (Hay-Jahans, 2019, p. 271-273).

## 3.4 Two-sample $t$-test

The two-sample $t$-test is a statistical test used to compare the means of two independent samples. This test is commonly employed when the variances of the two samples are not known. In a pooled $t$-test, the assumption is made that the variances of the two samples are equal. When conducting this test, the null hypothesis states that there is no significant difference between the means of the two populations, while the alternative hypothesis suggests a difference between the means of the populations under investigation. Let there be k populations, and $\mu_i$ and $\mu_j$ are the means of the populations. The null and the alternate hypothesis is as follows,

$$H_0 : \mu_i = \mu_j \quad \text{vs.} \quad H_1 : \mu_i \neq \mu_j \quad ,$$

To compare the means of multiple populations, we conduct pairwise $t$-tests between each pair of populations $(i, j)$, where $i$ and $j$ range from 1 to $k$ and $i \neq j$. This results in a total of $\frac{k(k-1)}{2}$ pairwise $t$-tests .

Let's consider $k$ populations, each with sample sizes $n_1, ..., n_k$ and corresponding sample variances $s_1^2, ..., s_k^2$. When performing a $t$-test between the $i$-th population with sample mean $\bar{y}_i$ and the $j$-th population with sample mean $\bar{y}_j$, we calculate the test statistic $t^*$ based on the null hypothesis where $t^* \sim t\,(n_1 + n_2 - 2)$. The formula for $t^*$ is as follows,

$$t^* = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2+...+(n_k-1)s_k^2}{n_1+n_2+...+n_m-m}} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \quad ,$$

The test statistic $t^*$ follows a $t$-distribution with degrees of freedom equal to the sum of the sample sizes minus the number of populations being compared. To determine if there is a significant difference between the means, we compare the absolute value of $t^*$ to a critical value obtained from the $t$-distribution table at a certain significance level $(\alpha)$. Moreover, the $p$-value represents the probability of obtaining a test statistic as extreme as or more extreme than $t^*$ under the null hypothesis. If the $p$-value is greater than the chosen significance level, it suggests that the observed difference in means could be due to random chance, and we do not reject the null hypothesis. In addition, if the $p$-value is smaller than the significance level, we reject the null hypothesis and conclude that there is a significant difference between the means of the populations being compared. In summary, the $t$-test helps assess whether the means of different populations differ significantly based on the calculated test statistic $t^*$ and the corresponding $p$-value compared to the chosen significance level (Rasch et al., 2020, p. 73).

## 3.5 Multiple comparisons problem

Multiple comparisons involve considering more than two decisions simultaneously based on the results of multiple tests. Assessing the associated risks is crucial. In this scenario, $k$ populations are labelled $P_1, P_2, ..., P_k$. Each population has independent random variables $(x_i)$ following a normal distribution. The normal distribution has an unknown mean $(\mu_i)$ and an unknown common variance $(\sigma^2)$. Specific methods are used when the variances are unequal. This setup is similar to the one-way ANOVA framework, where the populations represent the levels of a fixed factor $K$. We draw independent random samples $(X_i)$ of size $(n_i)$ from each population. Some methods assume equal sample

sizes, while others accommodate unequal sample sizes. We address various multiple comparisons (MC) problems in these situations (Rasch et al., 2020, p. 375-376).

## 3.6 Bonferroni correction method

Bonferroni adjustment allows for determining a significance cutoff by dividing the desired significance level, denoted as $\alpha$, by the total number of independent tests conducted, denoted as $m$. For instance, if a significance threshold of 0.05 is chosen and $m$ independent tests are performed, the null hypothesis is rejected if the resulting $p$-value is less than $\frac{\alpha}{m}$. The smaller the $p$-value, the more substantial the evidence for rejecting the null hypothesis and indicating significance. Comparing each $p$-value ($m\times$ $p$-value) to $\alpha$ is essential to assess the overall significance level. Alternatively, one can maintain a constant significance level and adjust the $p$-value by multiplying it by the number of independent tests ($m\times$ $p$-value $< \alpha$). It is important to note that the $p$-value is a probability and should not exceed 1. The Bonferroni correction effectively controls the Type-I error rate in any situation (Hay-Jahans, 2019, p. 274).

## 3.7 Tukey's procedure and Confidence interval

Tukey's procedure, also known as Tukey's Honestly Significant Difference (HSD), is a method used in statistical analysis to compare the means of multiple groups in pairwise fashion. It provides a way to construct simultaneous confidence intervals for the differences between all possible pairs of means. These confidence intervals help identify statistically significant differences between the groups.

The formula for constructing the Tukey confidence intervals takes into account the sample means ($\bar{y}$), standard deviation ($s$), sample sizes ($n$), and a critical value ($q_\alpha$) determined based on the desired significance level ($\alpha$). The intervals are defined as,

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

Where $j$ and $k$ represent the groups being compared. Alternatively, the test statistic $q^*_{jk}$ can be calculated as,

$$q^*_{jk} = \frac{\sqrt{2}\,(\bar{y}_j - \bar{y}_k)}{s\sqrt{1/n_j + 1/n_k}} \sim q(k, n-k) \quad ,$$

Which follows a distribution with $q(k, n-k)$ degrees of freedom and are used to calculate $p$-value $= P\left(q \geq q^*_{jk}\right)$. The $p$-value can be compared against the significance level ($\alpha$) to determine if the pairwise difference is statistically significant (Hay-Jahans, 2019, p. 276).

## 3.8 Levene's test

Levene's test is a statistical method used to assess whether the variances of multiple groups or samples are equal. It is particularly useful when the underlying distributions deviate from normality but have symmetric properties. The test compares the absolute deviations of the observations from their respective group means.

To conduct Levene's test, we consider $k$ groups, each with a sample size of $n_j$, and denote the $i$-th observation of the $j$-th sample as $x_{ij}$. We calculate the absolute deviations $d_{ij}$ as the absolute differences between $x_{ij}$ and the mean of the $j$-th sample, $\bar{x}_j$. The mean of all absolute deviations is denoted as $\bar{d}$, and the sample means, and variances of the absolute deviations are denoted as $\bar{d_j}$ and $s^2_{dj}$, respectively. The test statistic, denoted as $F^*$, is computed as:

$$F^* = \frac{\sum_{j=1}^{k} n_j \left(\bar{d_j} - \bar{d}\right)^2 / (k-1)}{\sum_{j=1}^{k} \left(n_j - 1\right) s^2_{d_j} / (n-k)} \quad ,$$

where $F^*$ follows an F-distribution with degrees of freedom $(k-1)$ and $(n-k)$. The $p$-value is calculated as $P(F \geq F^*)$, representing the probability of observing a test statistic as extreme or more extreme than $F^*$.

In summary, Levene's test examines whether the variances across multiple groups are equal. It compares the absolute deviations of the observations from their group means using a test statistic that follows an $F$-distribution. A higher $p$-value suggests insufficient evidence to reject the assumption of equal variances, while a lower $p$-value indicates significant differences in variances among the groups (Hay-Jahans, 2019, p. 247-248).

# 4 Statistical analysis

## 4.1 Descriptive analysis of data

The histogram plots in Figure 1 visually represent the distribution of birth weights within each smoking status category. The histograms reveal distinct patterns for each category. Category 0 exhibits a roughly symmetric distribution centred around 120 ounces, while Category 1 shows a slight left-skew with a concentration of lower birth weights. Categories 2 and 3 display symmetric distributions, with peaks around 120 and 125 ounces, respectively. Category 9 has a higher concentration of higher birth weights. These observations provide an initial understanding of the weight distribution within each smoking status category. These histograms visually represent the birth weight distributions within each category, enabling us to observe their differences or similarities.

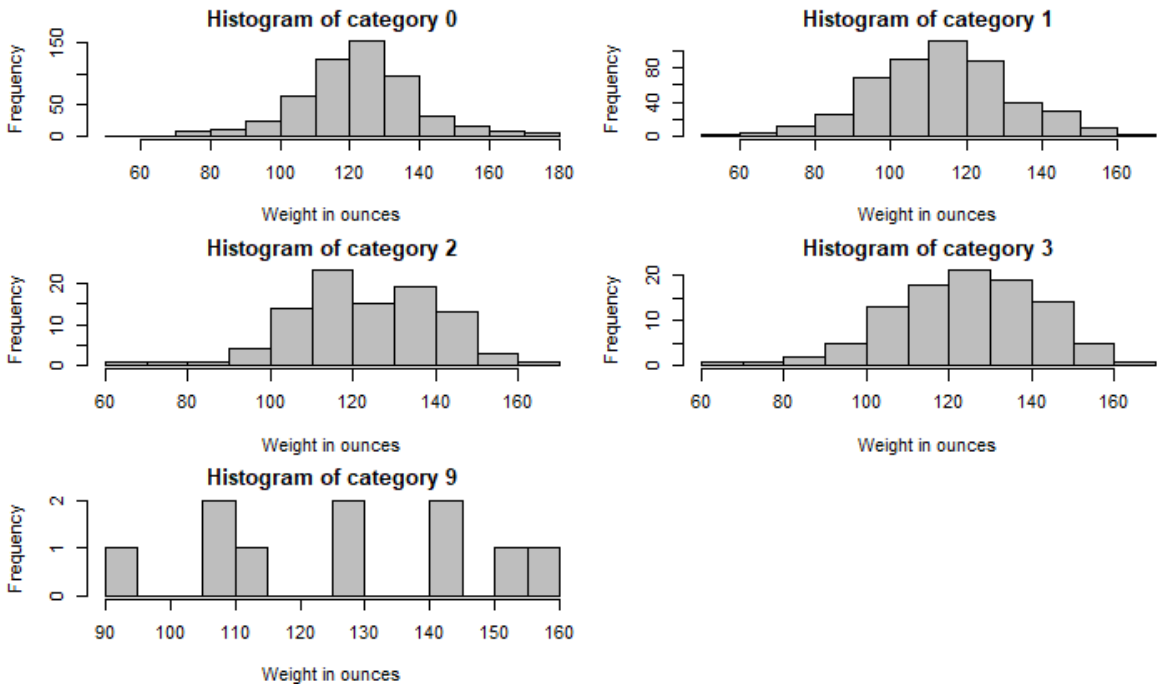


Figure 1: Frequency distribution of weights (in ounces)

Additionally, the summary statistics in Table 4 of the Appendix on page 18 provide numerical measures to describe each category's birth weight distribution. Regarding the continuous variable "wt" (Weight), the distribution and central tendency within each category can be analyzed. The average weight varies from 114.11 to 126.70 ounces

between categories 1 and 9. The median weight ranges from 115,00 ounces for Category 1 to 128,00 ounces for Category 9. The standard deviation (SD) quantifies the overall dispersion or variability of the weight data within each category. Category 0 values range from 17,06 ounces to Category 9 values of 21,81 ounces. The interquartile range (IQR) is the difference between the middle 50 percent of weight values within each category. It ranges between 18.50 ounces for Category 0 and 32.00 ounces for Category 9.

Figure 2 shows the Q-Q plots of the weights for the five different smoking categories. The $x$-axis represents the theoretical quantile, and the $y$-axis represents the sample quantile of the weights for each smoking category.

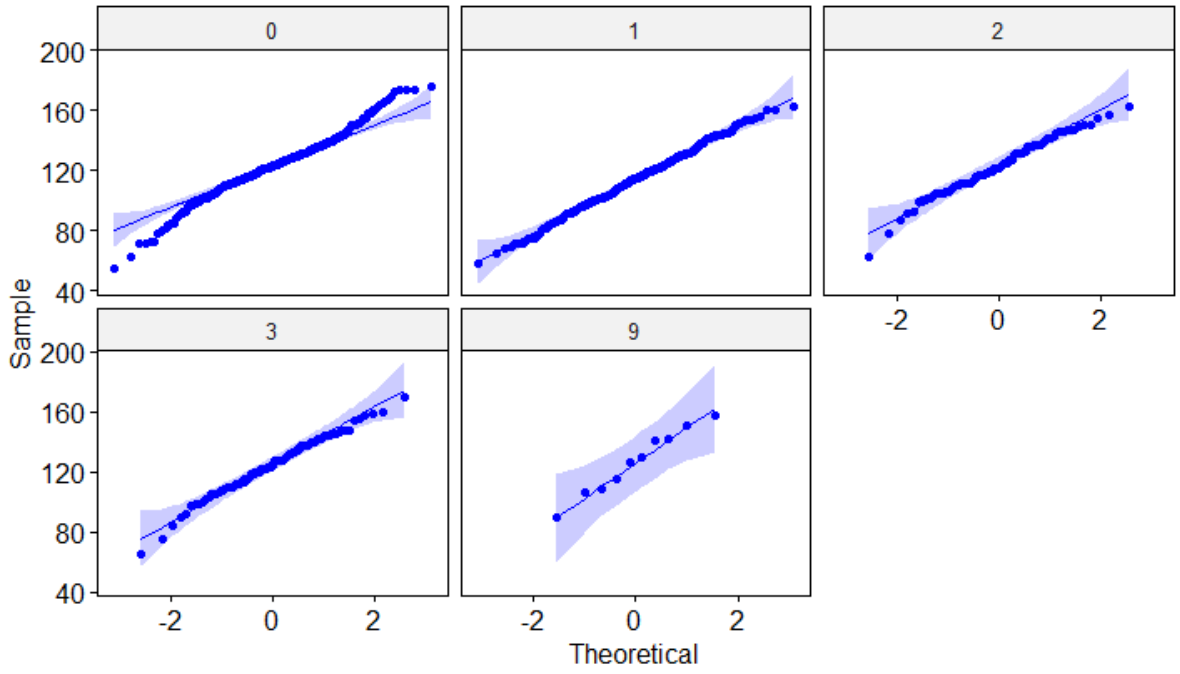


Figure 2: Q-Q plots of weights for different smoking categories

From the Q-Q plots, we can observe that the points for all categories approximately fall on a straight line, indicating that the weights are normally distributed within each smoking category. This suggests that the assumption of normality is reasonable for the weights in each smoking category. The Q-Q plots provide visual evidence of the normality assumption and support the use of parametric statistical tests that rely on this assumption when analyzing the weights across different smoking categories.

We fail to reject the null hypothesis using Levene's test since the $p$-value is greater than the significance level. This indicates that there is no evidence to suggest a statistically

significant difference in the variances of weights among the different smoking categories. There exists homogeneity of variance in the weight distribution across the five categories.

## 4.2 Global test

We apply one-way ANOVA test in this subsection. Here we compare the means of five smoking categories. The null hypothesis, denoted as $H_0$, posits that the average weight (measured in ounces) is equal across all five categories. Conversely, the alternative hypothesis, denoted as $H_1$, proposes that there are variations in the average weight (measured in ounces) among the five categories. In Table 1, the ANOVA provides important insights into the relationship between the "smoke" variable (representing different smoking categories) and the weight distribution. This Table 1 summarizes the results of a global test, specifically a Type-II test, which aims to determine if the "smoke" variable significantly impacts weight.

| Variable | DFn | DFd | F Value | $p$-value |
|---|---|---|---|---|
| Smoke | 4 | 1221 | 19.62 | $1.15 \times 10^{-15}$ |

Table 1: ANOVA test results.

The $F$ statistic, calculated as 19.62, and the extremely small $p$-value of $1.15 \times 10^{-15}$ (significantly less than the 0.05 threshold) indicate a strong statistical significance. This means that the average weight across the various smoking categories is different.

## 4.3 Two-sample $t$-test with adjustment methods and comparisons

The pairwise comparisons using $t$-tests with pooled standard deviation (SD) are conducted to examine the differences in birth weights (wt) between different smoking categories (smoke). Table 2 presents the $p$-values for each pairwise comparison. The $p$-values represent the probability of observing the observed differences in birth weights or more extreme differences under the null hypothesis that there is no difference in birth weights between the compared categories. In addition, the $p$-values reported in Table 2 are not adjusted for the number of comparisons made.

The pairwise comparisons using $t$-tests with pooled standard deviation (SD) are performed to assess the differences in birth weights (wt) between different smoking categories (smoke).

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 1 | $5.6\times10^{-15}$ | - | - | - |
| 2 | 0.91 | $6.4\times10^{-6}$ | - | - |
| 3 | 0.36 | $7\times10^{-8}$ | 0.54 | - |
| 9 | 0.50 | 0.026 | 0.538 | 0.724 |

Table 2: $p$-value of Pairwise comparisons (Without adjustment method).

Table 3 presents the $p$-values for each pairwise comparison after applying the Bonferroni adjustment to control for multiple comparisons. Table 3 displays the adjusted $p$-values obtained from pairwise comparisons using the Bonferroni adjustment method. In Table 3, a value of $5.6 \times 10^{-14}$ indicates that the $p$-value for the comparison between Category 1 and Category 0 is less than the chosen significance level (0.05). Therefore, the null hypothesis of no difference in birth weights between these categories is rejected. Similarly, a $p$-value of $6.4 \times 10^{-5}$ for the comparison between Category 2 and Category 1 is also less than the significance level, leading to the rejection of the null hypothesis. On the other hand, for comparisons where the $p$-value is 1.00 (e.g., Category 3 and Category 2), we fail to reject the null hypothesis as the $p$-value is greater than the significance level. These results suggest significant differences in birth weights between certain categories, while no significant differences are found in other pairs.

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 1 | $5.6\times10^{-14}$ | - | - | - |
| 2 | 1.00 | $6.4\times10^{-5}$ | - | - |
| 3 | 1.00 | $7\times10^{-7}$ | 1.00 | - |
| 9 | 1.00 | 0.26 | 1.00 | 1.00 |

Table 3: $p$-value of Pairwise comparisons (With Bonferroni adjustment method).

The results of Tukey's HSD analysis from Table 5 in the Appendix on page 18 reveal that significant pairwise differences in birth weights are found between Category 1 and Category 0 (p adj $< 0.05$), Category 2 and Category 1 (p adj $< 0.05$), and Category 3 and Category 1 (p adj $< 0.05$). These comparisons indicate that the birth weights in these pairs of categories are significantly different.

However, no significant differences are observed between other categories after adjusting for multiple comparisons using Tukey's HSD. It should be noted that the adjusted $p$-values are larger than the unadjusted $p$-values, reflecting the conservative nature of the adjustment to account for multiple testing.

We can observe some notable differences by comparing the results of the two correction methods, Bonferroni and Tukey's HSD, with the non-adjusted test. In the non-adjusted test, the $p$-values for several pairwise comparisons are below the significance level of 0.05, indicating significant differences in birth weights between certain categories. However, when applying the Bonferroni correction, most of the adjusted $p$-values become larger than 0.05, resulting in no significant differences between any pairs of categories. Similarly, the adjusted $p$-values obtained using Tukey's HSD are larger than 0.05 for most pairwise comparisons, indicating no significant differences after adjustment.

The reason for these differences is the need to control the overall Type-I error rate when conducting multiple hypothesis tests. The non-adjusted test does not account for the increased probability of obtaining false positives (Type-I errors) when performing multiple tests simultaneously. As a result, it may lead to an inflated number of significant findings and an increased chance of making incorrect conclusions.

In contrast, the Bonferroni correction is a conservative method that divides the significance level by the number of comparisons, making it more difficult to reject the null hypothesis and decreasing the likelihood of false positives. Tukey's HSD also adjusts the $p$-values to control the overall Type-I error rate while considering the sample sizes and variances of the compared groups.

Applying the Bonferroni correction and Tukey's HSD in this particular analysis led to a more stringent criterion for determining statistical significance. As a result, the number of significant differences is reduced, highlighting the importance of considering multiple testing corrections to avoid spurious findings and ensure reliable statistical inference.

Overall, comparing the results of the Bonferroni correction, Tukey's HSD, and the non-adjusted test provides a comprehensive understanding of the impact of multiple testing corrections on the interpretation of the data and helps to avoid drawing erroneous conclusions based on individual pairwise comparisons.

# 5 Summary

The statistical report aimed to investigate the relationship between maternal smoking and infant weight by analyzing data from the "Babies" dataset. The report employed various statistical techniques to address the research questions, including descriptive statis-

tics, global tests, Levene's test, pairwise comparisons, and adjustment methods such as the Bonferroni correction and Tukey's Honest Significant Difference (HSD) method.

The descriptive analysis revealed distinct patterns in the distribution of birth weights across different smoking status categories. The summary statistics provided numerical measures to describe the birth weight distributions, supporting the insights gained from the histograms. Q-Q plots demonstrated that the weights within each smoking category approximately followed a normal distribution, validating the assumption of normality for further analysis.

Levene's test showed no evidence of significant differences in the variances of weights among the smoking categories, indicating homogeneity of variance. The global test, conducted through ANOVA, indicated a strong statistical significance, suggesting that the average weight across the smoking categories differed significantly.

Pairwise comparisons using $t$-tests were performed to assess differences in birth weights between smoking categories. The unadjusted $p$-values revealed significant differences in birth weights between Category 1 and Category 0, Category 2 and Category 1, and Category 3 and Category 1. After applying the Bonferroni correction and Tukey's HSD method, the adjusted $p$-values showed no significant differences in birth weights between any pairs of categories except for the aforementioned comparisons.

Comparing the results of the adjusted tests with the non-adjusted test highlighted the importance of multiple testing corrections. The Bonferroni correction and Tukey's HSD provided more stringent criteria for determining statistical significance, reducing the number of significant differences and ensuring more reliable statistical inference.

In conclusion, the analysis revealed significant differences in birth weights between specific pairs of smoking categories, emphasizing the impact of maternal smoking on neonatal weight. The report emphasized the need to consider multiple testing corrections to avoid false findings and draw accurate conclusions based on the data. The findings contribute to understanding the relationship between maternal smoking and infant weight, providing insights for further research and potential interventions.

# Bibliography


Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. URL https://CRAN.R-project.org/package=gridExtra. R package version 2.3.

Berkeley Statistics. STAT LABS: Data, June 2023. URL https://www.stat.berkeley.edu/users/statlabs/labs.html. [Online; accessed 12th Jun. 2023].

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.

Christian Heumann, Michael Schomaker, and Shalabh. *Introduction to Statistics and Data Analysis With Exercises, Solutions and Applications in R*. Springer International Publishing, Cham, Switzerland, 2016. ISBN 978-3-319-46160-1. doi: 10.1007/978-3-319-46162-5.

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.

Paul Newbold, William L. Carlson, and Betty M. Thorne. *Statistics for Business and Economics*. Pearson Education Limited, Edinburgh Gate, Harlow, Essex CM20 2JE, England, 8th edition, 2013. ISBN 978-0-273-76706-0. URL http://www.pearson.com/uk.

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/.

Dieter Rasch, Rob Verdooren, and Jürgen Pilz. Applied statistics: Theory and problem solutions with r, 2020. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781119551584.

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.

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 | wt(weight) | 540.00 | 55.00 | 176.00 | 122.86 | 124.00 | 18.50 | 17.06 |
| 1 | wt(weight) | 481.00 | 58.00 | 163.00 | 114.11 | 115.00 | 24.00 | 17.97 |
| 2 | wt(weight) | 95.00 | 62.00 | 163.00 | 123.08 | 122.00 | 24.50 | 17.80 |
| 3 | wt(weight) | 100.00 | 65.00 | 170.00 | 124.63 | 124.50 | 26.00 | 18.57 |
| 9 | wt(weight) | 10.00 | 90.00 | 158.00 | 126.70 | 128.00 | 32.00 | 21.81 |

Table 4: Summary Statistics of Weight by Category (Measured in ounces).

| Comparison | 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 and Confidence interval for Tukey's adjustment method.