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: Vishesh Srivastava

Group number: 1

Group members: Arindam Pal, Jaimin Prashantkumar Oza

June 9, 2023

Contents

1 Introduction

Smoking by pregnant women causes severe issues for their babies, including fetal injury, premature birth, or low birth weight. In addition, these children may face many physical and mental problems during their lifetime (Stat Labs, 2002). Thus, exploring the relationship between maternal smoking and their newborns' weight is of great interest, including the effects of different smoking conditions. The data set used consists of 1236 observations and 23 variables. However, our analysis focuses on the variables *weight*, representing the babies' birth weight in ounces, and *smoke*, indicating the various categories of the smoking habits of mothers (Stat Labs, 2002).

Descriptive statistics summarize *weight* and *smoke* variables' distribution, while hypotheses test the relationship between maternal smoking categories and babies' birth weight. Before applying any statistical test, all underlying assumptions are checked. A global test, i.e., one-way analysis of variance checks if babies' birth weights differ across different smoking categories. Subsequently, a pairwise comparison is made between the resulting birth weights for different smoking categories using a two-sample *t*-test. These results are adjusted for multiple comparison problems using Bonferroni correction and Tukey's Honest Significant Difference with Tukey's confidence interval to measure uncertainty around observed differences. Finally, a comparison is made between the results of the non-adjusted and adjusted tests to identify potential biases.

Descriptive analysis shows that the babies' weight among different smoke categories ranges from 12.58 ounces for the mean and 4.81 ounces for the standard deviation. The global test observes a significant difference in babies' birth weights among different smoking categories. The pairwise comparison using a two-sample t-test with Bonferroni correction and Tukey's HSD shows significant birth weight differences among the pairs '*smokes now - never*,' '*until current pregnancy - smokes now*' and '*once did but not now - smokes now*.'

The second section provides a detailed description of the data set and its quality. The third section discusses the statistical methods used in the project, namely hypothesis testing, Levene's test, Kolmogorov-Smirnov test, one-way analysis of variance, pairwise two-sample *t*-test, Bonferroni correction and Tukey's Honest Significant Difference with confidence interval information. The fourth section discusses the analysis by interpreting the results using the statistical methods defined in section three. The final section summarizes the main findings with a brief explanation and outlook.

2 Problem Statement

2.1 Data Set Description

The data set used in this project contains information related to the newborns' weight and their mother's smoking status. The U.S. Government of Chicago collects this data via an observational study (Stat Labs, 2002). The data set contains 1236 observations and 23 variables, out of which only two are of interest in this project, namely *smoke* and *weight* representing the mothers' smoking status divided into five categories and babies' birth weight in ounces, respectively. The *smoke* is a discrete numeric variable that provides details about the different categories of smoking among the mothers, where a value of 0, 1, 2, 3, 9 corresponds to *never*, *smokes now*, *until current pregnancy*, *once did but not now*, and *unknown* respectively. The *weight* is a continuous numeric variable where a value of 999 corresponds to an unknown birth weight. Table 1 shows the count of each smoke category, with category *never* having the highest count and *unknown* having the lowest count.

Table 1: Count of data points based on different smoke categories

Smoking Category	never	smokes now	until current pregnancy	once did but not now	unknown
Count	544	482	99	101	10

The data set contains ten missing observations for the *weight* variable, out of which four were missing from category 0 (*never*), one from category 1 (*smokes now*), four from category 2 (*until current pregnancy*), and one from category 3 (*once did but not now*). Before the analysis, all these missing values are substituted with their respective group mean values of the variable *weight.* The *unknown* category from the *smoke* variable is removed before further analysis because of the unknown smoke type. Also, the count for these variables is less than one percent of the total data; thus, it will not affect our results. The *weight* variable with values '999' is also removed as those are unknown. The data is top-quality, as missing values are addressed before the analysis. In addition, the data set used is a subset of a much larger data set from the Child Health and Development Studies, including information on all pregnancies between 1960 and 1967 among women of the Kaiser Foundation Health Plan in Oakland, California, with a participation of more than 15,000 families, which further validates the data quality (Stat Labs, 2002).

2.2 Project Objectives

The content-related objectives include exploring the relationship between different maternal smoking categories and babies' birth weight. The project aims to determine whether there are observable differences in babies' birth weights among different categories of maternal smoking. The results provide insights into the impact of smoking during pregnancy on infant health. The main goal is to derive an association between maternal smoking and babies' birth weight through which females can adopt healthier behaviors during pregnancy.

The statistical objectives include a descriptive analysis to understand the data set variables, i.e., babies' birth weight and maternal smoking, calculating count, mean, variance, standard deviation, minimum and maximum values. The statistical assumptions are validated using Levene's for equality of variances and the Kolmogorov-Smirnov test for the normality of residuals. A global test is used to determine if there are differences in birth weights across different smoke categories. A pairwise comparison uses a two-sample t-test to identify specific birth weight differences between different smoke categories. To correct for multiple comparison problems, i.e., type I error, the Bonferroni method and Tukey's Honest Significant Difference with confidence interval information are used. At last, a comparison is made between the results from the non-adjusted and adjusted methods.

3 Statistical Methods

This section defines the statistical methods used in the analysis with their mathematical formulas and definition. All the analyses are performed using software R (Version 4.2.1, R Core Team, 2022), including the packages dplyr (Wickham, 2023) and car (Fox and Weisberg, 2019).

3.1 Hypothesis Testing

Hypothesis testing is a statistical method used to draw conclusions about the population based on sample data. It helps us make informed decisions and gain insights into the underlying characteristics of the population. It allows us to make inferences and identify patterns in the data when access to the entire population is impractical. Hypothesis testing is performed by formulating the null hypothesis (H_0), stating that the general population has no significant effect, variation, or association. In comparison, the alternate hypothesis (H_1) states a significant effect, a significant difference, or a significant relationship in the general population. By analyzing the sample data, we can determine whether to reject the H_0 or if we fail to reject the H_0 (Gravetter and Wallnau, 2015, p. 224-228).

3.1.1 The Critical Region and Type I, Type II Error

The critical region is a range of values or conditions defined in a hypothesis test. It is determined based on the chosen significance level (α), representing the maximum probability of rejecting the H_0 when true. If the test statistic falls within the critical region, it provides evidence against the H_0, leading to its rejection. The critical region is typically defined to capture extreme or unlikely values inconsistent with the H_0.

Type I error, also known as a false positive, occurs when the H_0 is rejected even though it is true. In hypothesis testing, it represents the probability of erroneously concluding that a specific intervention or factor is effective when, in reality, it has no impact. The α value for a hypothesis test is the probability of the occurrence of a Type I error.

Type II error, also known as a false negative, occurs when the investigator fails to reject H_0 even though it is false. In other words, a type II error means that the hypothesis fails to detect an actual intervention or factor. Type II error is denoted by β.

Table 3 on page 17 in the Appendix shows the possible results of a statistical hypothesis test (Gravetter and Wallnau, 2015, p. 230-238).

3.1.2 Test Statistic and p-Value

Hypothesis testing uses sample data to calculate the test statistic, which is crucial in decision-making. The critical value is obtained from distribution tables (e.g., t-table, z-table) by selecting the α and degree of freedom and locating the corresponding intersection. By comparing this critical value with the test statistic, one can evaluate the strength of evidence against the H_0 (Gravetter and Wallnau, 2015, p. 230).

The p-value is the probability assuming H_0 is true, indicating no significant difference among the variables. It quantifies the likelihood of obtaining the observed data point due to randomness (Dodge, 2008, p. 434-435). Given a predetermined threshold α, two scenarios can arise:

$$\text{If } p \leq \alpha \text{, reject } H_0 \text{, and if } p > \alpha \text{, fail to reject } H_0$$

3.2 Levene's Test for Equality of Variances

Levene's test is a statistical method used to assess the homogeneity of variances among multiple groups. It is essential because many statistical analyses, such as analysis of variance or t-test, rely on the assumption of equal variances across groups. The test compares variances within each group to determine if significant differences exist. The null hypothesis (H_0) claims equal variances across all groups, while the alternative hypothesis (H_1) suggests at least one group has a significantly different variance. It can be written as:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_m^2 \quad \text{and} \quad H_1 : \exists(i,j) \text{ such that } \sigma_i^2 \neq \sigma_j^2$$

Here $1 \leq (i,j) \leq m$, $i \neq j$, and m are the group or sample numbers.

The test statistic (L_{stat}) for Levene's test is calculated as:

$$L_{\text{stat}} = \frac{(N-m)\sum_{i=1}^{m} N_i(\bar{z}_{group} - \bar{z}_{\text{overall}})^2}{(m-1)\sum_{i=1}^{m}\sum_{j=1}^{N_i}(z_{ij} - \bar{z}_{group})^2}$$

Here N is the total sample size, N_i is the sample size of the i^{th} subgroup, $\bar{z}_{group}$ are the group means of the z_{ij}, $\bar{z}_{\text{overall}}$ is the overall mean of z_{ij}. The L_{stat} is approximately F-distributed with $m - 1$ and $N - m$ degrees of freedom (DF), and α is the chosen significance level. The critical value for rejecting the H_0 is given by:

$$L_{stat} > F_{1-\alpha,m-1,N-m}$$

Here $F_{\alpha,m-1,N-m}$ is the upper critical value of the F-distribution with (m - 1) and (N - m) DF at α. Levene's test uses an F-distribution table to obtain a p-value which can also be used to make the test decision after comparing to the chosen α value. If the p-value is less than or equal to α, the H_0 is rejected; otherwise, there is insufficient evidence to reject the H_0 (Jiang, 2022, p. 215-216).

3.3 Kolmogorov-Smirnov Test for Normality of Residuals

The one-sample Kolmogorov-Smirnov (KS) test helps determine if the residuals from a statistical model follow a normal distribution or significantly deviate from it. This is done by estimating how well the observed residuals fit the theoretical normal distribution. The null hypotheses (H_0) and alternate (H_1) hypotheses for the KS test are as follows:

$$H_0 : F(x) \sim N(\mu, \sigma^2) \quad \mathit{and} \;\; H_1 : F(x) \nsim N(\mu, \sigma^2)$$

Here $F(x)$ denotes the distribution of the residuals. $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. The H_0 claims the residuals follow a normal distribution, while H_1 claims the contrary. The KS test statistic (KS_{stat}) measures the maximum vertical distance between the cumulative distribution function (CDF) of the observed residuals and the CDF of the theoretical normal distribution. The KS_{stat} is formulated as follows:

$$KS_{stat} = \sup_x |F_{\text{obs}}(x) - F_{\text{theo}}(x)|$$

Here $F_{\text{obs}}(x)$ represents the empirical CDF of the observed residuals, and $F_{\text{theo}}(x)$ represents the CDF of the theoretical normal distribution. An empirical estimate is provided by the observed CDF, which is based on observed data. On the other hand, the theoretical CDF represents the distribution based on theoretical assumptions and is derived from a probability distribution model.

The H_0 is rejected if the associated p-value with KS_{stat} is less than α while if the p-value is greater than α, then our attempt to rule out H_0 fails (Dodge, 2008, p. 283-286).

3.4 One-Way Analysis of Variance

A one-way analysis of variance or ANOVA checks if significant differences exist between the mean values of two or more independent groups with only one explanatory variable. However, it acts as a global test and does not provide specific information about the groups for which mean values are significantly different if significant differences exist. The model can be formulated as follows:

$$y_{ij} = \mu_j + \varepsilon_{ij}$$

Here $i \in \{1, 2, \ldots, l\}$, $j \in \{1, 2, \ldots, n_l\}$, l is the number of groups levels in the data set, y_{ij} represents the observed value of the dependent variable for the i^{th} observation in the j^{th} group, μ_j represents the overall mean of the j^{th} group or treatment and ϵ_{ij} represents the random error associated with the i^{th} observation in the j^{th} group (Dodge, 2008, p. 9-10).

3.4.1 Assumptions and Hypothesis

The following three assumptions must be satisfied before applying the ANOVA test:

1. Data is sampled randomly and independently.
2. Variances of each sample are equal.
3. Residuals follow a normal distribution.

The null hypothesis (H_0) suggests that the mean values are the same for all groups, while the alternate hypothesis (H_1) suggests that at least one of the mean values differs significantly:

$$H_0\text{: } \mu_1 = \mu_2 = \ldots = \mu_n \text{ and } H_1\text{: } \exists i, j \text{ such that } \mu_i \neq \mu_j$$

Here μ_i and μ_j $(1 \leq (i, j) \leq n$ and $i \neq j)$ are the mean values in the i^{th} and j^{th} sample respectively (Black, 2010, p. 407-409).

3.4.2 Calculating F-Test Statistic

The primary concept behind conducting a one-way ANOVA is to partition the total variation or Total Sum of Squares (TSS) of the data into two components: the within-group variance or Sum of Squares Within (SSW) and the between-group variance or Sum of Squares Between (SSB).

Suppose, there are l groups, with $n_1, n_2, ..., n_l$ observations in each group. The total number of observations is $n = \sum_{i=1}^{l} n_i$. The sample means of the groups are denoted by $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_l$, and the overall mean is denoted by $\bar{y}$. Further, y_{ij} represents the individual data points in the group i and observation j. Then the mathematical formulas are written as:

$$\text{TSS} = \sum_{i=1}^{l}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2;\ \ \text{SSB} = \sum_{i=1}^{l} n_i \cdot (\bar{y}_i - \bar{y})^2;\ \ \text{SSW} = \sum_{i=1}^{l}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

The Mean Square Treatment (MST) is the variation among the group means, obtained by dividing TSS by the degree of freedom (DF), i.e., n - 1. The mean square (MSA) is the variation among groups, obtained by dividing the SSB by its DF, i.e., l - 1. The mean square within (MSW) is the residual variation within each group, obtained by dividing the SSW by its DF, i.e., $n - l$.

$$\text{MST} = \frac{\text{TSS}}{n-1};\ \text{MSA} = \frac{\text{SSB}}{l-1};\ \text{MSW} = \frac{\text{SSW}}{n-l}$$

Furthermore, the test statistic (F_{stat}) is given by:

$$F_{\text{stat}} = \frac{\text{MSA}}{\text{MSW}} = \frac{\sum_{i=1}^{l} n_i \cdot (\bar{y}_i - \bar{y})^2/(l-1)}{\sum_{i=1}^{l}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(n-l)}$$

Once a significance level (α) is chosen, the decision rule is to reject the H_0 if $F_{\text{stat}} > F(\alpha, l-1, n-l)$. Calculating the p-value corresponding to F-statistic uses F-distribution. If the p-value is less than or equal to α, reject the H_0 concluding a significant difference between the means exists. In contrast, a p-value greater than α results in the failure to reject the H_0; this suggests insufficient evidence for a significant difference among the means (Black, 2010, p. 408-409).

3.5 Pairwise Two-Sample t-Test

The independent two-sample t-test is used in hypothesis testing to compare the mean value for two independent groups. Two samples, x_i and x_j $(i \neq j)$, are derived from a population with a normal distribution and the same variance. In addition, the observations in each sample are also independent of each other. The null hypothesis (H_0) and the alternate (H_1) hypotheses are as follows:

$$H_0 : \mu_i - \mu_j = 0,\ H_1 : \mu_i - \mu_j \neq 0.$$

The test statistic ($t_{statistic}$), which follows a t-distribution for the t-test under H_0, can be specified as:

$$t_{statistic} = \frac{\overline{x}_i - \overline{x}_j}{s_{pooled}\sqrt{\frac{1}{m_i} + \frac{1}{m_j}}}$$

Here $\overline{x}_i$ and $\overline{x}_j$ are the sample means of two different groups and m_i and m_j represents the size of samples with $1 \leq (i, j) \leq n$ and $i \neq j$, s_{pooled} is the pooled sample standard deviation of population standard deviation σ. The s^2_{pooled} represents the pooled sample variance and can be written as:

$$s^2_{pooled} = \frac{\sum_{i=1}^{n}(m_i - 1)s_i^2}{\sum_{j=1}^{n} m_j - n}$$

Here n is the total number of groups or samples, m_i $(i = 1, 2, \ldots, n)$ represents the sample size of n groups. The denominator in s^2_{pooled} represents the degrees of freedom (DF), whereas s_i^2 $(i = 1, 2, \ldots, n)$ represents the sample variances of n respective groups. The H_0 is rejected if the $t_{statistic}$ falls outside the critical region defined by the critical values $-t_{1-\alpha/2;\mathrm{DF}}$ and $t_{\alpha/2;\mathrm{DF}}$ that are obtained from the t-distribution with a pre-defined significance level (α). Conversely, the H_0 cannot be rejected if the computed t-value falls within the range $-t_{1-\alpha/2;\mathrm{DF}} \leq t_{statistic} \leq t_{\alpha/2;\mathrm{DF}}$. The DF is given as; $\mathrm{DF} = \sum_{j=1}^{n} m_j - n$, where m_j $(j = 1, 2, ..., n)$ represents the sample sizes of j^{th} group.

Calculating the p-value from the t-distribution table and then comparing it with α can also make this decision: it is given by p-value $= 2P(t_{\mathrm{DF}} > |t|)$ under the assumption that H_0 is true, where t is the test statistic value and t_{DF} represents the t-distributed random variable with $\mathrm{DF} = \sum_{j=1}^{n} m_j - n$ (Noether, 2012, p. 128-133).

3.6 Multiple Testing Problem

There are often situations where various hypothesis tests are needed for each comparison, for example, a pairwise two-sample t-test or ANOVA. However, the higher the statistical tests being conducted, the higher the probability of obtaining statistically significant results by chance alone. This may lead to erroneous conclusions and the identification of false relationships among variables.

Consider a set of n null hypotheses denoted as $H_{01}, \ldots, H_{0n}$, concerning Type I errors where the family-wise error rate (FWER) is the probability of observing at least one Type I error. The FWER can be defined as:

$$\mathrm{FWER} = \Pr(V \geq 1).$$

Here V is the count of Type I errors or false positives. By rejecting any null hypothesis with a p-value below the significance level (α), the resulting FWER is given by:

$$\mathrm{FWER} = 1 - \Pr(V = 0) = 1 - \Pr\left(\bigcap_{k=1}^{n} \{\text{no false rejection of } H_{0j}\}\right)$$

The FWER can be expressed as follows, assuming that each of the n null hypotheses is true and that each of the n tests are independent:

$$\text{FWER}(\alpha) = 1 - (1 - \alpha)^n$$

(James et al., 2021, p. 560-562).

3.6.1 Bonferroni Correction

The Bonferroni correction method maintains a preferred family-wise error(FWER) rate by altering each test's significance level (α). This method adjusts the α for multiple comparisons by dividing it with the number of independent tests denoted as n. This method is not applied if $p * n > 1$, where p is the original p-value. The corrected significance level is denoted as $\alpha_{corrected}$ is given as:

$$\alpha_{corrected} = \frac{\alpha}{n}$$

The probability of wrongly rejecting any null hypothesis in the Bonferroni method is less than or equal to α. Let Z be an event of falsely rejecting at least one null hypothesis and Z_i be the event of falsely rejecting the i^{th} null hypothesis. The $P(Z)$ can be written as:

$$P(Z) = P\left(\bigcup_{i=1}^{n} Z_i\right) \leq \sum_{i=1}^{n} P(Z_i) = n.$$

The Bonferroni-corrected p-value ($p_{i,\text{Bonf}}$) for each test is obtained by multiplying the real p-value (p) with the number of tests (n): $p_{i,\text{Bonf}} = n \cdot p_i$, where $1 \leq i \leq n$. The Bonferroni-corrected rejection region for each test is denoted as R_i and is obtained by comparing the $p_{i,\text{Bonf}}$ to the $\alpha_{corrected}$ (Wasserman, 2005, p. 165-166). It is given as follows:

$$R_i = \{H_{0i} \text{ is rejected if } p_{i,\text{Bonf}} < \alpha_{\text{corrected}}\}$$

3.6.2 Tukey's Honest Significant Difference and Confidence Interval

To address multiple testing problems, John Tukey introduced intervals based on the sample range, not individual differences. These intervals are known as Tukey's Honestly Significant Difference (HSD) intervals which are only suitable for balanced designs having several observations to be the same for each factor level. It incorporates an adjustment for the sample size allowing for a sensible interval even when little unbalanced designs are present. Tukey's HSD, denoted as T_{HSD} can be given as:

$$T_{HSD} = z_{\alpha,c,\nu} \cdot \sqrt{\frac{\text{MSE}}{n}}$$

Here $z_{\alpha,c,\nu}$ is the critical value from the studentized range distribution with α as significance level and $\nu = N$ - c, N is the total number of observations, and c is the total number of groups, MSE is the mean square error which can be calculated as the sum of squares within (as defined in ANOVA) dividing by the corresponding degree of freedom, i.e., ν, n is the number of observations present in each group. The p-values are obtained by making a comparison of critical values to studentized range statistics (q-statistics), which are calculated by dividing mean differences by standard error.

Tukey's confidence interval estimates the true difference between the two groups. It measures uncertainty around the estimated difference and is computed based on the variability among the sample means. The Tukey confidence interval is written as follows, where considering the negative sign in the formula gives the lower value and the positive will give the upper value:

$$\text{CI} = (\bar{y}_i - \bar{y}_j) \pm T_{HSD},$$

Here $\bar{y}_i$ and $\bar{y}_j$ are the means of the i^{th} and j^{th} $(i \neq j)$ groups respectively and its subtraction gives the mean difference between two groups (Black, 2010, p. 418-421).

4 Statistical Analyses

In this section, the statistical analysis and hypothesis testing are performed using the methods defined in Section 3 after checking all underlying assumptions. As a preprocessing step, all missing values of the variable *weight* are substituted with the group mean of the respective category. In addition, unknown values for the variable *weight* and *smoke* are also removed. A detailed description of the results from descriptive statistics and the statistical test is given. The analyses denote the null and alternate hypotheses by H_0 and H_1, respectively, and a significance level (α) uses a value of 0.05.

4.1 Descriptive Statistics and Distribution Analysis

Descriptive statistics uses two variables *smoke* and *weight*. The *smoke* variable is discrete and divided into four categories: *never*, *smokes now*, *until current pregnancy*, and *once did but not now*. In contrast, the variable *weight* is continuous and represents the babies' birth weight in ounces. Table 4 on page 17 in the Appendix summarizes the statistics for the four categories of the *smoke* and *weight* variables. The *never* category contains the highest value for the count and maximum, i.e., 544 and 176 ounces, respectively, while the lowest value for standard deviation (SD) and minimum, i.e., 17.00 and 55 ounces, respectively. The *smokes now* category contains 482 observations and has the lowest mean value of babies' birth weight of 114.11 ounces. The next category, i.e., *until current pregnancy*, contains the lowest number of observations, i.e., 99, and a mean value of babies' birth weight of 124.08 ounces. The last category, *once did but not now* contains 101 observations and the highest mean, SD, and minimum value, i.e., 124.63, 18.48, and 65 ounces, respectively. Overall, the table gives in-depth insights into the data distribution across different smoke categories, with the mean values ranging from 114.12 to 124.63 ounces. In comparison, the SD ranges from 17.00 to 18.48 ounces.

4.2 Global Test for Differences in Birth Weights

Before applying the ANOVA global test, all three assumptions are checked. The first assumption of random and independent sampling is valid since the data points in each group are independent of those in the other group, and a random sample obtains the observations present in each group. The second assumption of equal variances among groups uses Levene's Test with H_0, stating that all groups have equal variances. At the same time, H_1 suggests that at least one of the variances is unequal. Levene's test results in a p-value of 0.10, as shown in Table 5 on page 17 in the Appendix, which is greater than the α suggesting no strong evidence against H_0. Validating the last assumption to check if the residuals follow normal distribution is based on the Kolmogorov-Smirnov (KS) test. The H_0 of the KS test states that the data is drawn from a normally distributed population, while the H_1 states the opposite. The test results in KS_{stat} value of 0.032 and p-value of 0.17. Since the p-value exceeds α, we fail to reject H_0.

As our data set satisfies all the assumptions, the ANOVA test can be applied. The H_0 and H_1 for ANOVA are written as follows where μ represents the mean weight in the respective category:

$$H_0\text{: } \mu_1 = \mu_2 = \mu_3 = \mu_4 \text{ and } H_1\text{: } \exists i, j \ (i \neq j) \text{ such that } \mu_i \neq \mu_j, \text{ where } 1 \leq (i, j) \leq 4$$

The results of the one-way ANOVA test, as shown in Table 2, include information on the degree of freedom, the sum of squared error, the mean squared error, the F-value, and the p-value. Since the obtained p-value is smaller than the α, the data has statistical evidence to reject the H_0 stating all means are equal. Thus, at least one group with a significantly different mean value exists.

Table 2: Results of one-way ANOVA test conducted at a significance level (α) of 0.05

	DF	SSB	MST	F value	p-Value
Smoke	3	24080	8027	26.10	2.31×10^{-16}
Residuals	1222	375862	308		

4.3 Pairwise Differences Between the Resulting Birth Weights

The global test shows that at least one pair of the mean weights of babies differ across categories. A pairwise t-test is used to get exact information on which pair of categories the birth weight differs significantly. The same assumptions apply to a one-way ANOVA and a pairwise two-sample t-test. As the same data set as that of ANOVA is utilized, these assumptions are met. The H_0 and H_1 are formulated as:

$$H_0\text{: } \mu_i = \mu_j \text{ and } H_1\text{: } \mu_i \neq \mu_j$$

Here μ_i and μ_j ($i \neq j$ and $1 \leq (i, j) \leq 4$) are the mean birth weights in the i^{th} and j^{th} categories, respectively.

The p-values obtained from the pairwise t-test at α value of 0.05 are shown in Table 6 on page 18 in the Appendix. A significant difference in the birth weight is observed for pairs *'smokes now - never'*, *'until current pregnancy - smokes now'* and *''once did but not now - smokes now'* as the p-value obtained is smaller than α; this suggests a high impact of smoking during pregnancy on fetal development. Furthermore, the study did not reveal any significant difference in the babies' birth weights when comparing the pairs *'until current pregnancy - never'*, *'once did but not now - never'*, and *'once did but not now - until current pregnancy'*, as the p-value is obtained is greater than α, resulting in non-rejection of the H_0. Thus, these pairs do not show any significant difference in the mean birth weight of babies. However, knowing the potential possibility for Type II errors and interpreting the results cautiously is important.

4.4 Comparison of Correction Methods with Non-Adjusted Test

Since multiple hypothesis testing increases the risk of type I error, taking care of this issue is essential. Bonferroni and Tukey's Honest Significant Difference (HSD) are the two adjustment methods for this problem. Table 7 on page 18 in the Appendix summarizes the results for six comparisons of smoke categories before and after the correction with the information of p-value and the decision to reject the H_0 or not, where 'Y' tells the H_0 is rejected. Simultaneously, the H_0 is not rejected when' N' occurs.

The pairwise comparison of categories *'smokes now - never'*, *'until current pregnancy - smokes now'* and *'once did but not now - smokes now'* shows a p-value smaller than α before and after the correction. This results in a rejection of the H_0, indicating a significant difference in mean birth weights between both categories. On the other hand, while examining the pairwise comparison of *'until current pregnancy - never'*, *'once did but not now - never'*, *'once did but not now - until current pregnancy'*, the non-adjusted and the adjusted methods suggest the same conclusions, this results in a failure to reject the H_0, indicating no significant difference in mean birth weights between both categories.

While Bonferroni and Tukey's HSD provides a more rigorous control for the multiple comparisons, both correction methods resulted in the same conclusions as the non-adjusted ones. This also suggests that the non-adjusted results are not only due to chance. However, we must employ the correction methods to avoid any possibility of Type I error in our results.

Table 8 on page 18 in the Appendix shows Tukey's Confidence Interval providing additional information on the pairwise differences in the babies' birth weights between different smoke categories. It contains three columns; the first denotes the mean difference, and the second and third denotes the lower and upper value of the confidence interval. The *'once did but not now - smokes now'* comparison shows the highest positive value for the mean difference of 10.52 ounces. This shows that the babies born to females who smoked once but do not smoke now to females who currently smoke have, on average, a higher birth weight of babies. The confidence interval is from 5.58 ounces to 15.46 ounces, providing a range under which the true population difference lies. On the other hand, The *'smoke now - never'* comparison shows the lowest negative mean difference value of -8.75 ounces. This shows that the babies born to females with a current smoking status have, on average, a lower birth weight than those born to non-smokers and a confidence interval of -11.58 ounces to -5.93 ounces.

5 Summary

This project studied the relationship between different maternal smoking categories during pregnancy and their babies' birth weights in a subset of the Child Health and Development Studies data. The analysis used two of the 23 independent variables provided in the data set; the first was *smoke* denoting different smoke categories, and the second was *weight* denoting the babies' birth weight in ounces. The Introductory Case Studies course instructors, SoSe-2023, provided this data set.

The descriptive analysis considered the count of data points for the variable *smoke.* The results show the highest value for the *never* category and the lowest for the *until current pregnancy* category. In contrast, central tendency measures such as mean, variance, and standard deviation with information on maximum and minimum were examined for the variable *weight.* The results suggested that the average value of mean birth weight can differ by up to 10.52 ounces in the examined data set. In contrast, on average, the variability of the data points can differ by up to 1.48 ounces from the mean.

The global test checked whether or not significant differences exist in the babies' birth weights among different smoking categories. The results revealed that at least one pair exists in different smoke categories with an unequal mean. The analysis followed a pairwise two-sample t-test to compare the babies' birth weight between different smoke categories. It was adjusted using Bonferroni and Tukey's Honest Significant Difference (HSD) with information on Tukey's confidence interval using α value of 0.05. Both the adjusted and non-adjusted tests reached the same conclusion, rejecting the null hypothesis for the pairs; *'smoke now - never'*, *'until current pregnancy - smokes now'*, and *'once did but not now - smokes now'* stating that a significant difference exists between the mean values of these pairs. In contrast, we fail to reject the assumption that mean values are equal for the rest pairs of smoke categories.

In conclusion, the analysis revealed significant differences in birth weights among the different maternal smoking categories during pregnancy. At the same time, the pairwise comparisons provided a deeper insight into the differences between the different smoke categories. The Bonferroni correction and Tukey's HSD were used to control Type I error for multiple comparison problems, ensuring the results' reliability. However, it is crucial to interpret these results cautiously and avoid generalizing beyond the study's scope. Further investigations could explore the influence of additional aspects that may have affected the babies' birth weights and consider possible confounding variables.

Bibliography

Black K. (2010): Business Statistics: For Contemporary Decision Making, Sixth Edition, John Wiley & Sons, Inc., Hoboken, NJ.

Dodge, Y. (2008): *The Concise Encyclopedia of Statistics*. Springer Science+Business Media, LLC. Available at: `https://doi.org/10.1007/978-0-387-32833-1`

Fox, J., and Weisberg, S. (2019). car:*An R Companion to Applied Regression*, Third edition. Sage, Thousand Oaks, CA. `https://socialsciences.mcmaster.ca/jfox/Books/Companion/`.

Gravetter, F. J., and Wallnau, L. B. (2015). *Statistics for the Behavioral Sciences* (10th ed.). Boston, MA: Cengage Learning.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer Texts in Statistics. Los Angeles, CA: University of Southern California.

Jiang, J. (2022): *Applied Medical Statistics*. Hoboken, NJ: John Wiley & Sons, Inc.

Noether, G. E. (2012): *Introduction to Statistics: The Nonparametric Way*. Springer Science & Business Media.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. url: https://www.R-project.org/.

Stat Labs: Concepts, Models, and Applications (2002): University of California, Berkeley. URL: `https://www.stat.berkeley.edu/users/statlabs/labs.html` (visited on 18th May 2023).

Wasserman, L. (2005): *All of Statistics: A Concise Course in Statistical Inference*. Springer. Available at: `https://doi.org/10.1007/978-0-387-21736-9`

Wickham, H. (2023). dplyr: A Grammar of Data Manipulation. R package version 1.0.8. URL: `https://cran.r-project.org/package=dplyr`.

Appendix

Table 3: Potential results in a statistical hypothesis test

Decision	H_0 True	H_0 False
Fail to reject H_0	Correct	Type II error
Reject H_0	Type I error	Correct

Table 4: Summary statistics of four smoke categories and babies' birth weight

Category	Count	Mean	Variance	Std. Dev.	Min	Max
never	544	122.86	288.91	17.00	55	176
smokes now	482	114.11	322.35	17.95	58	163
until current pregnancy	99	123.08	304.03	17.44	62	163
once did but not now	101	124.63	341.41	18.48	65	170

Table 5: Levene's test results for the assumption of equal variances of each sample

	DF	L_{stat}	P-value
group	3	2.06	0.10

Table 6: Results from pairwise t-test conducted at a significance level (α) of 0.05

Pairwise Comparison	never	smokes now	until current pregnancy
smokes now	3.40×10^{-15}	-	-
until current pregnancy	0.91	3.91×10^{-6}	-
once did but not now	0.35	5.10×10^{-8}	0.53

Table 7: Summary of test results for the six pairs of smoke categories before pairwise t-test and after the Bonferroni correction and Tukey's HSD at α of 0.05

Pairs	Before Correction		After Bonferroni		After Tukey's HSD	
	P-value	Reject H_0-Y/N	P-value	Reject H_0-Y/N	P-value	Reject H_0-Y/N
smokes now - never	3.40×10^{-15}	Y	2.00×10^{-14}	Y	0.00	Y
until current pregnancy - never	0.91	N	1.00	N	1.00	N
once did but not now - never	0.35	N	1.00	N	0.79	N
until current pregnancy - smokes now	3.90×10^{-6}	Y	2.30×10^{-5}	Y	2.31×10^{-5}	Y
once did but not now - smokes now	5.10×10^{-8}	Y	3.10×10^{-7}	Y	3.00×10^{-7}	Y
once did but not now - until current pregnancy	0.53	N	1.00	N	0.92	N

Table 8: Tukey's confidence intervals for pairwise differences

Pairs	mean difference	lower	upper
smokes now - never	-8.75	-11.58	-5.93
until current pregnancy - never	0.22	-4.71	5.15
once did but not now - never	1.77	-3.12	6.66
until current pregnancy - smokes now	8.98	3.99	13.95
once did but not now - smokes now	10.52	5.58	15.46
once did but not now - until current pregnancy	1.55	-4.84	7.93