TU Dortmund

Introductory Case Studies

Project II: Comparison Of Multiple Distributions

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

Author: Bushra Tariq Kiyani

Group number: 2

Group members: Sarmistha Bhattacharyya , Jatin Rattan , Vikas Kumar

June 8, 2023

Contents

1 Introduction

Baby weight is an important metric that provides valuable information about a newborn's health and well-being. A baby's birth weight is considered a crucial indicator of its growth and development during the pregnancy. Both low and high birth weights are associated with certain health risks and can be influenced by various factors such as diet, growth, and the intrauterine environment. Maternal smoking during the pregnancy can have adverse effects on the baby's birth weight, potentially increasing the risk of low birth weight babies for mothers who smoke. Investigating the association between maternal smoking and a newborn's weight is crucial to understanding the effects of smoking on a baby's health and development. Identifying high-risk groups and understanding the mechanisms by which smoking affects birth weight can help design targeted intervention programs to promote child health (Reeves and Bernstein, 2008).

This project aims to compare the multiple distributions of a sample taken from the dataset (Stat Labs, UC Berkeley). The main goal of this project is to find the relationship between maternal smoking and babies' weight, and whether different conditions of maternal smoking lead to changes in the weight of different groups of neonates. The dataset contains the babies' birth weight in ounces and the mother's smoking history into five categories (never, smokes now, until current pregnancy, once did not now, unknown). We want to analyze birth weight variations between the categories. To do this, we first look at the frequency distributions of smoke through counts and babies' birth weights using histograms. Boxplots are used to study the variations within and between all categories. After that, we check the distribution of the data using Q-Q Plots. Then, to test whether weights differ between the categories, we perform ANOVA. Finally, to observe the pairwise differences in weights of all categories, we conduct two-sample tests for all pairs and then compare the test results after adjusting with the Bonferroni's correction and Tukey's Honest Significant Difference (HSD)/Tukey-Kramer test.

The second section describes the structure and quality of the dataset in more detail. Furthermore, we state the goals of the project in the second section. The third section explains the hypothesis testing, one-way ANOVA, Bonferroni, and Tukey's Honest Significant Difference methods. It also describes the graphical representations used in statistical analysis (histograms, boxplots, Q-Q plots). The fourth section focuses on the application of these methods and the interpretation of the graphs and results. Finally, the fifth section summarizes the most important findings and discusses possible further data analyses.

2 Problem statement

2.1 Data set and data quality

This report deals with the analysis of a small sample of the dataset "Maternal Smoking and Infant Health" taken from Stat Labs, an online collection of case studies that integrate the theory of statistics with the practice of statistics. It is a part of the University of California, Berkeley's Statistics Department and aims to teach statistics through real-world applications and case studies. (Stat Labs, UC Berkeley).

The dataset consists of 1236 observations and 24 variables. The project involves only the baby's weight and the mother's smoking status. Therefore, only three variables will be kept in the dataset: id, wt, and smoke. The description of these variables, based on the provided 'Babies data description' from the lecturers in Moodle, is as follows:

Variable	Type	Description
id	Numeric	Identification number.
wt	Numeric	Birth weight of the child in ounces (999 unknown).
smoke	Numeric	Does mother smoke? 0 = never, 1 = smokes now, 2 = until current pregnancy, 3 = once did, not now, 9 = unknown

Table 1: Data Description

There are 10 observations that are duplicated in the dataset. Additionally, there are 10 observations where the smoke category is 9, indicating the unknown smoking status of the mother. Moreover, there are 10 observations where weight values are missing. The quality of data in Stat Labs is generally good, as the labs are designed to provide real-world applications and case studies.

2.2 Project objectives

This project aims to analyze whether the variation in baby weight among the four categories of smoking is significant. We begin by examining the frequency distribution of weight in all categories using histograms. The interpretation is based on the mean and variance of the distributions. Next, we assess the assumptions necessary for applying ANOVA and t-tests to analyze the variability in the mean baby weight across all categories. Box plots are used to examine the homogeneity of variances in each category, and

Q-Q plots are used to assess the assumption of normally distributed data. The ANOVA test is then conducted to compare the mean weight differences between categories.

To investigate pairwise differences between baby weights, we perform pairwise t-tests. The results are interpreted by comparing the p-value at a significance level of 0.05. Subsequently, to address the issue of multiple testing, we adjust the results using Bonferroni and Tukey's Honest Significant Difference methods. The adjusted results are interpreted and compared with the unadjusted ones to account for multiple testing.

3 Statistical methods

3.1 Confidence Intervals

A confidence interval is a method of estimating the likely range of values in which a population parameter falls. It is used to account for the uncertainty associated with the estimation process. A certain confidence level, such as 95 percent, is chosen to determine the width of the interval. A confidence interval has two components: an upper bound and a lower bound. These limits define the range of values expected to contain the true population parameter at the chosen confidence level. To calculate the confidence interval, the sample estimate is used as the midpoint, and the margin of error is added to and subtracted from that estimate. The margin of error is determined by multiplying the standard error (SE) of the estimate by a value corresponding to the chosen confidence level, denoted as T. The confidence interval is calculated as (Mood et al., 1973, p. 373):

$$\text{Sample Estimate} \pm T \times SE$$

3.2 Hypothesis Testing

Hypothesis testing is an inferential statistical method of analyzing whether a hypothesis can be accepted or rejected. Two types of hypotheses are formulated: the null hypothesis H_0, and the alternative hypothesis H_1. The null and the alternative hypotheses are set up before performing the hypothesis testing. The goal is to collect evidence that either supports rejecting the null hypothesis in favor of the alternative hypothesis or fail to reject the null hypothesis due to insufficient evidence. The **Null Hypothesis** H_0 is a concise mathematical statement that shows that there is no difference between some

characteristics of the data or groups being compared. The null hypothesis is denoted as "H_0" and is formulated in a way that can be tested statistically. The **Alternative Hypothesis** H_1 is the complementary statement to the null hypothesis, denoted as "H_A" or "H_1". An alternative hypothesis suggests that there is a significant relationship or difference between variables or groups (Mood et al., 1973, p. 402-405).

Test Statistic: A test statistic (e.g., t-value) is a random variable, calculated from sample data, which is subject to sampling variability. The distribution of the test statistic is determined by the sampling distribution under the null hypothesis. The distribution of the test statistic helps us quantify the evidence against the null hypothesis and make decisions in hypothesis testing (Mood et al., 1973, p. 419).

Level of Significance (α) and p-Value: The significance level α represents the required level of evidence to reject the null hypothesis. Typically, α is set to 5%(0.05). A lower α value indicates a need for stronger evidence to reject the null hypothesis. The p-value indicates the statistical significance of the obtained test results and also signifies the probability of committing a type I error in rejecting or not rejecting the null hypothesis. Its value always falls within the range of $[0, 1]$. The p-value is compared to the significance level α, and the null hypothesis is rejected if the p-value is smaller than α (Mood et al., 1973, p. 402-403).

Types of Error and Power of the test: In hypothesis testing, there are two types of errors: Type I errors (False Positive) and Type II errors (False Negative), both associated with incorrect conclusions about the null hypothesis. A Type I error occurs if we mistakenly reject the null hypothesis (H_0) when it is true. The probability of committing a Type I error is denoted as "α" and is set as the significance level of the test. Conversely, a Type II error occurs when we fail to reject the null hypothesis (H_0) even though it is false. The probability of making a Type II error is denoted as "β," and $1 - \beta$ represents the power of the test, which is the probability of correctly rejecting the null hypothesis when it is false (Mood et al., 1973, p. 405).

3.3 One-Way ANOVA Test

ANOVA (Analysis of Variance) is a statistical method that determines whether the means of multiple groups significantly differ from each other. It examines the variance levels between and within these groups. There are many variations of ANOVA but the main two types are one-way ANOVA and two-way ANOVA. One-way ANOVA is used

to assess the impact of a single independent variable on the dependent variable (Rasch et al., 2020, p. 108).

ANOVA Assumptions: Three primary assumptions about the data must be met before applying ANOVA: Normality: The responses for each sample group are taken from a normal population distribution. Equal variances: The variances of the populations that the samples come from are equal. Independence: All samples are drawn independently of each other (Rasch et al., 2020, p. 108-109).

Grand mean: In ANOVA calculations there are two types of the mean: separate sample means ($\mu_1, \mu_2, ..., \mu_k$) and the grand mean (μ). The grand mean is the mean of all sample means. We'll denote empirical grand mean by $\bar{x}_G$ and sample means as $\bar{x}_1, ..., \bar{x}_k$, then the grand mean is defined as (Rasch et al., 2020, p. 108):

$$\bar{x}_G := \frac{1}{k} \sum_{i=1}^{k} \bar{x}_i$$

Hypothesis: In ANOVA the null and the alternate hypothesis are:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k \text{ (means of all groups are equal)}$$

$$H_1 : \text{at least one } \mu_i \text{ is different from the other means; } i = 1, ..., k$$

The null hypothesis is valid when the sample means don't have significant differences while the alternate hypothesis is accepted when at least one of the sample means is different from the rest (Rasch et al., 2020, p. 114).

Variability Between Groups: It is the variation between the distributions of the individual groups. To calculate between-group variability the difference between the individual sample means and the grand mean is calculated. If the samples deviate greatly from each other, the difference between the individual mean and the overall mean would therefore also be significant. For sample sizes $n_1, ..., n_k$, the sum of squares for between-groups variability $\mathrm{SS}_{\mathrm{Between}}$, is calculated as (Rasch et al., 2020, p. 108-112):

$$\mathrm{SS}_{\mathrm{Between}} = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x}_G)^2$$

To calculate between-group mean squared deviation divide the $\mathrm{SS}_{\mathrm{Between}}$ by degrees of freedom (df_{B}): $\mathrm{k} - 1$ where k is the number of groups. Between-group mean squared

deviation is denoted as $\text{MS}_{\text{Between}}$ and calculated as (Rasch et al., 2020, p. 108-112):

$$\text{MS}_{\text{Between}} = \frac{\text{SS}_{\text{Between}}}{\text{k} - 1}$$

Variability Within Groups: As the variability of each sample is increased, their distributions overlap and they become part of a large population. Because values within each group are not the same this causes this variation. So the variance between individual points in each sample is calculated separately and referred to as within-group variability:

$$\text{SS}_{\text{Within}} = \sum_{i=1}^{K} \sum_{j=1}^{n_j} (x_{ij} - \bar{x}_i)^2$$

With degrees of freedom (df_{W}): N - k (N is the total number of observations), within-group mean squared deviation $\text{MS}_{\text{Within}}$ is calculated as (Rasch et al., 2020, p. 108-112):

$$\text{MS}_{\text{Within}} = \frac{\text{SS}_{\text{Within}}}{\text{N} - \text{k}}$$

Total variability can be calculated by adding both between and within groups variances.

F-Statistic: F-statistic (F) is the ratio of the between-group variance to the within-group variance. The formula for the F-statistic is:

$$\text{F} = \frac{\text{MS}_{\text{Between}}}{\text{MS}_{\text{Within}}}$$

The F-statistic, under H_0, follows a continuous probability distribution known as the F-distribution. The shape of the F-distribution is determined by the degrees of freedom associated with the numerator (dfB) and denominator (dfW) (Rasch et al., 2020, p. 108).

Decision rule in ANOVA: After determining the degrees of freedom associated with the F-statistic (df_{B}, df_{W}), the F-distribution table is consulted or statistical software is used to find the critical value of F at a given significance level (α) and degrees of freedom (df_{B}, df_{W}). The critical value represents the threshold beyond which the null hypothesis is rejected. Then calculated F-statistic is compared to the critical value. If the F-statistic is greater than the critical value, the null hypothesis is rejected. Otherwise, the test fails to reject the null hypothesis and concludes that there is not enough evidence to suggest significant differences between the means (Rasch et al., 2020, p. 108).

Problems with One-Way ANOVA: If the overall p-value from the ANOVA table is less than a certain level of significance, then we have enough evidence that at least one

group's mean is different from the other. However, this does not tell us which groups differ from each other. It just tells us that not all group means are equal (Rasch et al., 2020, p. 108-114).

3.4 t-Test

A t-test is a statistical test, used to compare the means of two groups. It is used in hypothesis testing to determine whether the means of the two groups are different from each other or not. It assumes that data is independent, taken from a normal distribution. A two-tailed t-test is conducted to determine if the two populations are distinct from one another, whereas a one-tailed t-test is performed to find whether one population mean is greater or lesser than the other (Christopher, 2019, p. 251-253).

Hypothesis: Consider two different samples from a normal distribution with expectations μ_1, μ_2 and constant variances. In the t-test the null and alternate hypotheses are formulated as (Christopher, 2019, p. 255):

$$H_0 : \mu_1 = \mu_2 \quad \text{and} \quad H_1 : \mu_1 \neq \mu_2$$

Performing a t-test: The t-test calculates the difference between two group means by dividing the difference in means by the pooled standard error::

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Here, t represents the t-statistic, $\bar{x}_1$ and $\bar{x}_2$ are the means of the compared groups, s^2 denotes the pooled standard error, and n_1 and n_2 indicate the sample sizes of each group. A large absolute value of t indicates a significant difference between the group means. The calculated t-value is compared to critical values in a table (e.g., Student's t-table). If the t-value exceeds the critical value the null hypothesis is rejected, indicating that the two groups are significantly different (Christopher, 2019, p. 273-274).

3.5 Multiple t-Tests

To compare the means of more than two groups multiple t-tests can be performed to check all pairwise differences. But the problem with this approach is each one of those

independent tests has an α level and α level is a type I error rate. The type I error compounds with each test. This alpha (α) inflation is caused by performing repeated statistical tests on the same data. Which will lead to a higher probability of making a Type I error and lower confidence (Christopher, 2019, p. 268).

Family-Wise Error Rate (FWER/FWE): The family-wise error rate represents the probability of committing at least one Type I error. When conducting m independent comparisons with a significance level of α for each test, the family-wise error rate is calculated as:

$$\tilde{\alpha} = 1 - (1 - \alpha)^m$$

In multiple t-testing, the probability of obtaining a significant result purely by chance continues to rise (Matsunaga, 2007).

3.6 The Bonferroni Correction

The Bonferroni correction fixes the significance level at $\frac{\alpha}{m}$. The Bonferroni correction is slightly conservative. If p_i is the p-value of hypothesis H_i, the Bonferroni correction rejects the null hypothesis if

$$p_i \leq \frac{\alpha}{m}$$

thereby controlling the FWER at $\leq \alpha$. Alternatively, we can compare each $m * p_i$ against joint significance level α. If the p-value becomes greater than 1, then its value is adjusted to 1 (Christopher, 2019, p. 274).

3.7 Tukey's HSD/Tukey-Kramer Test

Tukey's Honest Significant Difference, is a statistical test used to identify which specific pairs of means are different after an ANOVA test has found significant results. The HSD method compares pairwise means of groups while controlling the overall Type I error rate and is useful when dealing with multiple groups, where the goal is to determine which specific group means differ significantly from each other. **Performing HSD test:** After getting significant results from the ANOVA test, calculate the HSD test statistic (T_{HSD}) -when sample sizes are equal- as follows (Hochberg and Tamhane, 1987, p. 80):

$$T_{HSD} = q \times \sqrt{\frac{MS_{Within}}{n}}$$

T_{HSD} has studentized range distribution. In the formula, q is the critical value for this distribution based on the values of α (significance level), k (the number of groups), and df_W (degrees of freedom within the group). MS_{Within} is the mean square error within and n is the sample size. To deal with unequal sample sizes within each group the **Tukey-Kramer** method is used which is an extension of Tukey's HSD. The test statistic T_{TK} for unequal sample sizes is (Shiraishi et al., 2019, p. 2):

$$T_{TK} = q \times \sqrt{\frac{MS_{Within}}{n_i + n_j}} \quad i = 1, ..., k; j = 1, ..., k$$

Where n_i is the sample size in the i^{th} group and n_j is the sample size in the j^{th} group. **Decision rule:** Conduct pairwise comparisons and compare calculated test statistic (T_{HSD}/T_{TK}) to the differences between each pair of group means. If the absolute difference between the two means is greater than the T_{HSD}/T_{TK} value, it suggests that there is a significant difference between those groups. In the case of unequal sample sizes, we check: $|\bar{x}_i - \bar{x}_j| > T_{TK}$ (Shiraishi et al., 2019, p. 2).

Confidence Intervals: Confidence intervals are calculated by adding and subtracting the margin of error from the mean difference. In case of unequal sample sizes:

$$|\bar{x}_i - \bar{x}_j| \pm q \times \sqrt{\frac{MS_{Within}}{n_i + n_j}}$$

The confidence intervals indicate the range within which the true mean difference between the groups is likely to fall with a specified level of confidence. If the confidence interval includes zero, it suggests that there is no statistically significant difference between the corresponding groups. Otherwise, it indicates a statistically significant difference between the groups (Shiraishi et al., 2019, p. 3).

3.8 Quantile-Quantile (Q-Q) Plot

A Q-Q plot is a graphical tool that is used to compare two probability distributions. It is a scatterplot that is created by plotting quantiles of the distributions to be compared, against one another. The points form a roughly straight line ($y = x$) if both sets of quantiles came from the same distribution. It is a visual check, not a proof, we can see at-a-glance if the assumption of normality is plausible or not. **Normal Probability Q-Q Plot** gives a way for comparing the empirical quantiles of sample data against the

theoretical quantile of a normal distribution. Normal probability Q-Q plot takes sample data, sort it in ascending order, and then plots the empirical quantiles vs theoretical quantiles from $\mathcal{N}(\bar{x}, s^2)$ where $\bar{x}$ and s^2 are the sample mean and sample variance, respectively. If the points seem to fall about a straight line it can be assumed that the sample data came from a normally distributed population (Christopher, 2019, p. 147).

4 Statistical analysis

This section presents a descriptive analysis of the dataset using statistical measures and plots described in the previous section. For calculation of all statistical measures and graphical representations R software (R Core Team, 2022) Version, 4.2.1 is used with additional packages **ggpubr** (Kassambara, 2022a), **dplyr** (Wickham et al., 2022), **ggplot2** (Wickham, 2016), **gridExtra** (Auguie, 2017), **rstatix** (Kassambara, 2022b).

One instance of the 10 duplicated ids (1091, 2621, 6360, 6510, 7045, 7112, 7441, 8107, 8253, 8716) has the weight values, while in the other instance, the weight values are missing. We have retained the instances where the weight is not missing. Furthermore, 10 observations with the smoke category labeled as 9 (unknown) have been removed from the dataset. After the preprocessing steps, the dataset contains 1216 observations.

4.1 Frequency Distribution of Variables

Table 2 provides a summary of the descriptive statistics for the data. The number of observations differs across each smoke category. The count column indicates the sample size (no of observations) in each category. The categories "never" and "smoke now" have larger sample sizes with 540 and 481 observations, respectively, while the categories "until current pregnancy" and "once did, not now" have smaller sample sizes with 95 and 100 observations, respectively. The frequency distribution of weight in all categories is depicted in Figure 1 in the appendix. The x-axis represents the class intervals of the baby's weight, while the y-axis displays the frequency of each interval. The frequency distributions exhibit a nearly symmetrical pattern. In terms of central tendency, the mean and median values in all categories are quite similar, except for the "never" category where the median (124) slightly surpasses the mean (122.86). Notably, the mean and median weight in the "never", "until current pregnancy", and "once did, not now" categories show minimal variation among these categories. However, in the "smokes

Smoke Category	Count	Median	Mean	SD	Var	Min	Max
never (0)	540	124.0	122.86	17.06	291.05	55	176
smokes now (1)	481	115.0	114.11	17.97	323.02	58	163
until current pregnancy (2)	95	122.0	123.08	17.80	316.97	62	163
once did, not now (3)	100	124.5	124.63	18.57	344.86	65	170

Table 2: Summary Statistics

now" category, both the mean (114.11) and median (115) weights are notably lower compared to the other categories. This suggests that, on average, babies in the "smokes now" category tend to have lower weights compared to those in the other categories.

Furthermore, there is a considerably high variance particularly in the "once did, not now" category, which displays the highest variance of (344.86). Conversely, the "never" category exhibits the lowest variance (291.05). The "never" category exhibits two extreme outliers. The first outlier, identified by id 7722, has the lowest weight of 55 ounces among all categories. While, the second outlier, identified by the id 6760, has the highest weight of 176 ounces among all categories. These outliers represent exceptional cases that deviate from the typical weight distribution in the "never" category.

Outliers are also observed in the "smokes now," "until current pregnancy," and "once did, not now" categories. In the "smokes now" category, id 7544 has the lowest weight of 58 ounces, while id 7581 has the highest weight of 163 ounces. In the "until current pregnancy" category, the minimum weight is 62 ounces for id 7334, and the maximum weight is also 62 ounces for id 8122. Lastly, in the "once did, not now" category, id 7884 has the lowest weight of 65 ounces, while id 6660 has the highest weight of 170 ounces.

4.2 Verifying the Assumptions

As discussed in the previous section, in order to perform ANOVA and t-tests, we need to verify certain assumptions. In this section, we'll discuss if our data is fulfilling these assumptions.

Normality: We want to check whether the sample data is normally distributed or not. Figure 3 in the appendix shows Normal probability Q-Q plots of all four categories. We observe that data points do not follow the reference line well, mostly on the edges, especially in the "never" category. Nevertheless, we assume that our sample data is normally distributed. To be more precise we can use different tests (e.g. The Shapiro-Wilk test) to check the data normality.

Independence: As data (Stat Labs, UC Berkeley) was collected for educational purposes, randomly, we assume independence.

Variance Homogeneity: The third assumption is constant variance across all distributions. Figure 2 in the appendix shows box plots of babies' weights in all categories. All boxes' widths are almost similar except for the "never" category. But it's also not that much small. So overall there's not much difference in box plot widths. The whiskers also have almost similar lengths across the smoke categories, which suggests that the variances are comparable. Hence we assume variances are relatively similar. To be more accurate a test can be performed to verify variance homogeneity.

4.3 One-Way ANOVA: Testing Significance of babies' weight differences across the four smoke categories

As the three assumptions needed to perform ANOVA are fulfilled, now to test the significance of differences in the babies' weight across all categories, an ANOVA test is performed. Specifying μ as the mean weight of the categories and significance level $\alpha = 0.05$. The null and alternate hypotheses are formulated as follows:

$$H_0 : \mu_{never} = \mu_{smokesNow} = \mu_{untilCurrentPregnancy} = \mu_{onceDidNotNow}$$

$$H_1 : \text{at least one } \mu \text{ is different from the other means}$$

Table 3 shows the output of the ANOVA test conducted in R. In the table, the smoke is

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
smoke	3	23932	7977	25.72	3.91e-16 ***
Residuals	1212	375862	310		

Table 3: ANOVA Output

showing values between groups. The degree of freedom (DF) is $(k - 1 = 3)$, the sum of the square between categories (Sum Sq) is 23932 and the Mean square between categories is $\left(\frac{23932}{3} = 7977.3\right)$. Residuals show the values within groups. DF is $(N - k = 1212)$, the sum of the square within is 375862 while the Mean square is $\left(\frac{375862}{1212} = 310\right)$. The F value is then calculated. The p-value (Pr(>F)) is less than the significance level $\alpha = 0.05$, so the null hypothesis that the mean weight across all the categories is equal, is rejected.

4.4 Multiple t-tests

In this section, multiple t-test results are discussed. The assumptions required to perform a t-test have already been fulfilled. All possible pairs of smoke categories are as follows:

"never_smokesNow", "never_untilCurrentPregnancy", "never_onceDidNotNow",

"smokesNow_untilCurrentPregnancy", "smokesNow_onceDidNotNow",

"untilCurrentPregnancy_onceDidNotNow"

For the first pair, the hypotheses are formulated as follows: t-test1:

$$H_0 : \mu_{never} = \mu_{smokesNow} \quad \text{and} \quad H_1 : \mu_{never} \neq \mu_{smokesNow}$$

Likewise, we formulate hypotheses for all pairs and specify significance level $\alpha = 0.05$. After running the t-test separately on each pair in R, the results are given in Table 4.

	Categories pair	p-values	Reject Yes/No
0-1	never_smokesNow	<0.01	Yes
0-2	never_untilCurrentPregnancy	0.91	No
0-3	never_onceDidNotNow	0.91	No
1-2	smokesNow_untilCurrentPregnancy	<0.01	Yes
1-3	smokesNow_onceDidNotNow	<0.01	Yes
2-3	untilCurrentPregnancy_onceDidNotNow	0.55	No

Table 4: Multiple t-Test Results

The p-values for the three pairs "0-1", "1-2", and "1-3" are less than $\alpha = 0.05$, so the null hypotheses for these pairs are rejected. This shows that there is a significant difference in the mean weight between these pairs. While for the other three pairs p-value is greater than the significance level ($\alpha = 0.05$), the null hypotheses cannot be rejected.

4.5 The Bonferroni correction

To deal with the family-wise error rate we use Bonferroni's correction method. The p-values from multiple t-tests are adjusted using Bonferroni's method in R and the result is displayed in Table 5.

The p-values are adjusted and have increased compared to the multiple t-tests but the final result of rejecting/not rejecting the null hypotheses is the same as multiple t-tests.

	pair	adj p-values	Reject Yes/No
0-1	never_smokesNow	<0.01	Yes
0-2	never_untilCurrentPregnancy	1.00	No
0-3	never_onceDidNotNow	1.00	No
1-2	smokesNow_untilCurrentPregnancy	<0.01	Yes
1-3	smokesNow_onceDidNotNow	<0.01	Yes
2-3	untilCurrentPregnancy_onceDidNotNow	1.00	No

Table 5: Bonferroni's correction results

4.6 Tukey-Kramer test

As mentioned previously, the Tukey-Kramer test is an extension of Tukey's HSD method that enables pairwise comparisons between group means when the sample sizes are unequal across the categories. In this case, since the smoke categories have different sample sizes, the Tukey-Kramer test is performed instead of Tukey's HSD. According to the R documentation (R Core Team, 2021) for TukeyHSD(), the function can also handle unequal sample sizes. The results of the Tukey-Kramer test are presented in Table 6.

The adjusted p-values have been computed, and for the pair "0-3", it has decreased compared to the p-value in multiple t-tests. However, for the other pairs, the adjusted p-values have increased. Despite these adjustments, the final results of rejecting or not rejecting the null hypotheses remain the same as in the multiple t-tests.

	pair	adj p-values	Reject Yes/No
0-1	never_smokesNow	<0.01	Yes
0-2	never_untilCurrentPregnancy	1.00	No
0-3	never_onceDidNotNow	0.79	No
1-2	smokesNow_untilCurrentPregnancy	<0.01	Yes
1-3	smokesNow_onceDidNotNow	<0.01	Yes
2-3	untilCurrentPregnancy_onceDidNotNow	0.93	No

Table 6: Tukey-Kramer test results

Table 7 in the appendix presents Tukey-Kramer's confidence intervals. For the differences in means between the pairs "0-2", "0-3", and "2-3" the confidence intervals contain zero, suggesting that the differences between these means are not statistically significant. Conversely, the confidence intervals for the pairs "0-1", "1-2", and "1-3" do not include zero, indicating a statistically significant difference between the means of these pairs. The widest confidence interval [-4.94,8.04] is observed for the pair "2-3" while the confidence interval [-11.59,-5.91] for the pair "0-1" is the narrowest among all the pairs.

4.7 Summary

In this project, we conducted a comparison of multiple distributions using a dataset that included babies' weight at birth (in ounces) across four categories of maternal smoking status: "never," "smokes now," "until current pregnancy," and "once did, not now." The dataset was obtained from the Stat Labs website (Stat Labs, UC Berkeley), and our objective was to analyze the weight in each category. The main goal was to determine whether there were significant differences in weight among all categories. To achieve this, we performed ANOVA followed by multiple t-tests to explore pairwise differences. Additionally, we used two adjustment methods for multiple testing: Bonferroni correction, and the Tukey-Kramer test and compared their results with each other and with the multiple t-test outcomes.

The summary statistics revealed variations in sample sizes among the categories. Categories "never" and "once did, not now" exhibited the highest mean and median weights, but the sample sizes in these categories were 540 and 100, respectively. Conversely, "smokes now" displayed the lowest mean and median weight. The category "once did, not now" had the highest standard deviation of 18.57. The frequency distribution plots were generally symmetrical.

Prior to conducting the test, we verified the assumptions required for ANOVA and t-test. Data independence was assumed based on random data collection. Normality was assessed using Q-Q plots and homogeneity of variance was confirmed by weight boxplots analysis. The ANOVA yielded a significant p-value lower than $\alpha = 0.05$, indicating a notable difference among the mean weights of all categories. However, the ANOVA did not provide information about pairwise differences in weight across the categories.

To determine pairwise differences, multiple pairwise t-tests were conducted, yielding p-values less than α for three pairs: "0-1", "1-2", and "1-3". However, for the pairs "0-2", "0-3", and "2-3" p-values exceeded α, indicating that the null hypotheses cannot be rejected for these pairs. Multiple t-tests are susceptible to family-wise error rate, which were addressed by using adjustment methods. Following the application of Bonferroni's Correction and the Tukey-Kramer test, the p-values overall increased, with Bonferroni's Correction showing a larger increment compared to the Tukey-Kramer test. Nevertheless, both methods yielded the same result for the rejection or non-rejection of the null hypotheses as the multiple t-tests. In summary, for all pairs involving the "smokes now" category, the test results revealed a significant difference in mean weight, indicating noticeably lower weights in this category compared to the others.

Bibliography

Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. R package version 2.3.

Hay-Jahans Christopher. *Introduction to the theory of statistics*. Taylor Francis Group, 2019. 1st Edition.

Yosef Hochberg and Ajit C. Tamhane. *Multiple Comparison Procedures*. John Wiley Sons, Inc., 1987. ISBN 9780470316672. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316672. 1st Edition.

Alboukadel Kassambara. *ggpubr: ggplot2' Based Publication Ready Plots*, 2022a. URL https://rpkgs.datanovia.com/ggpubr/. R package version 3.4.

Alboukadel Kassambara. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*, 2022b. URL https://rpkgs.datanovia.com/rstatix/. R package version 3.3.

Masaki Matsunaga. Familywise error in multiple comparisons: Disentangling a knot through a critique of o'keefe's arguments against alpha adjustment. *Communication Methods and Measures*, 1(4):243–265, 2007. doi: 10.1080/19312450701641409. URL https://doi.org/10.1080/19312450701641409.

Alexander McFarlane Mood, Franklin A Graybill, and Duane C Boes. *Introduction to the theory of statistics*. McGraw-Hill, 1973. 3rd Edition.

R Core Team. *TukeyHSD function documentation*, 2021. URL https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/TukeyHSD. R package version 4.0.5.

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*. WILEY, 2020. 1st Edition.

Shane Reeves and Ira Bernstein. Effects of maternal tobacco-smoke exposure on fetal growth and neonatal size. 2008. URL https://ncbi.nlm.nih.gov/pmc/articles/PMC2770192/.

Taka-aki Shiraishi, Hiroshi Sugiura, and Shin-ichi Matsuda. *Pairwise Multiple Comparisons Theory and Computation*. Springer International Publishing, 2019. URL https://link.springer.com/book/10.1007/978-981-15-0066-4. 1st Edition.

Stat Labs, UC Berkeley. Maternal Smoking and Infant Health. URL https://www.stat.berkeley.edu/users/statlabs/labs.html. Accessed: 2023-05-30.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*, 2016. URL https://ggplot2.tidyverse.org.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2022. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

Appendix

A Additional figures

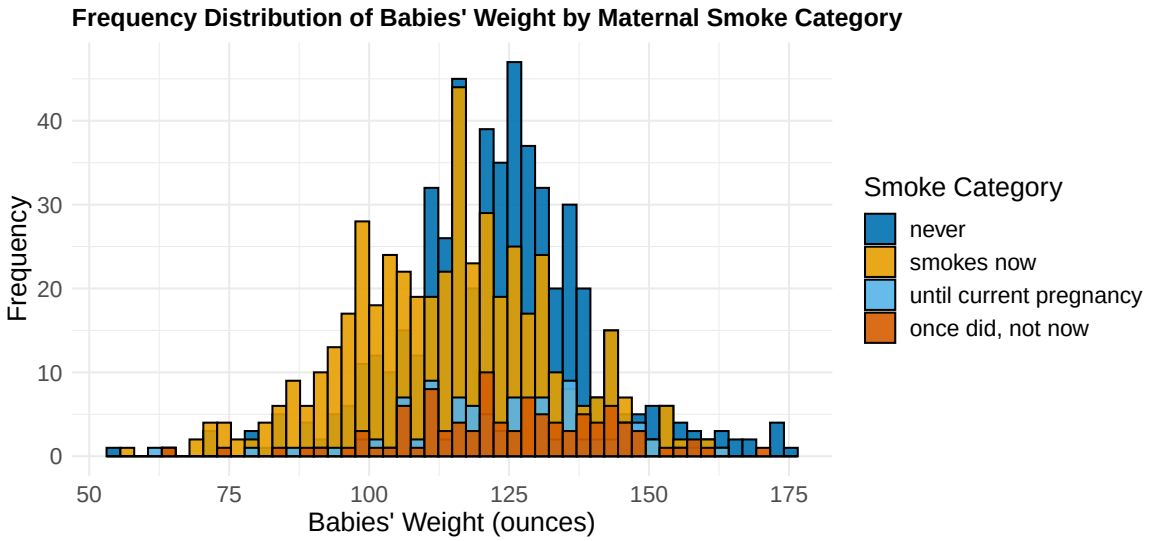

Figure 1: Histograms of frequency distributions of weight in all smoke categories

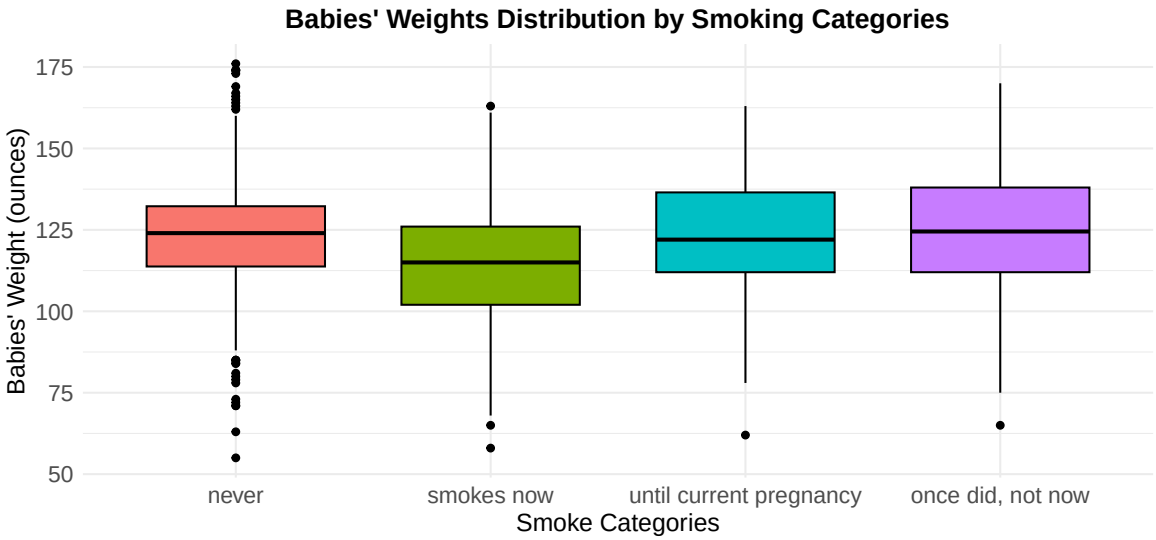

Figure 2: Box plots of babies' weight in all smoke categories

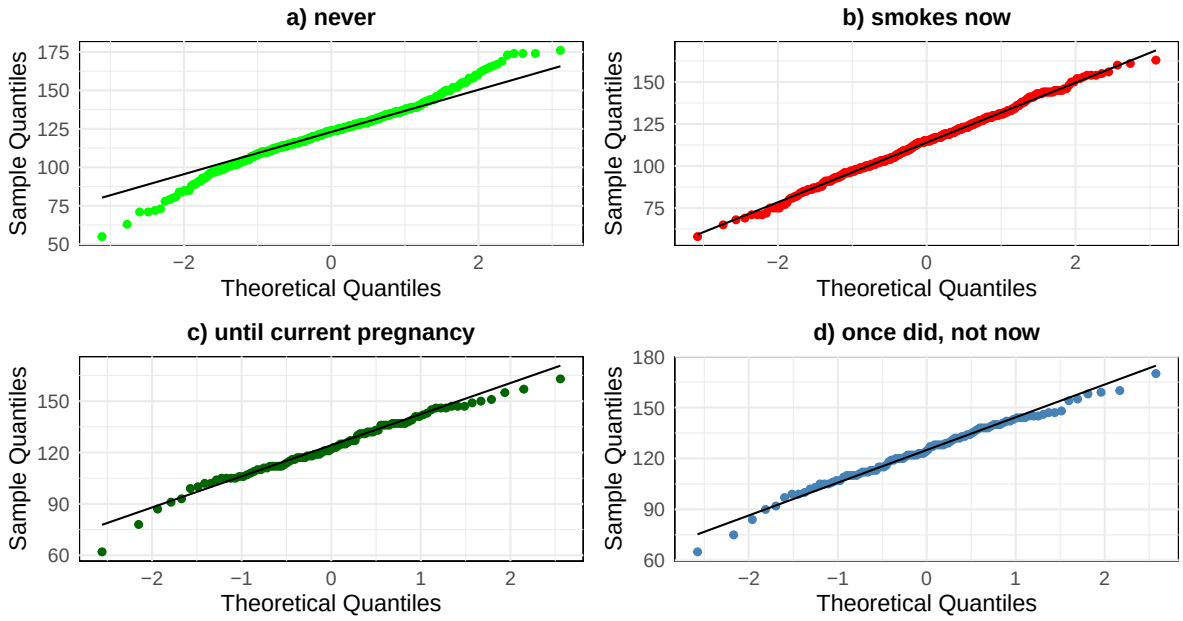

Figure 3: Normal Q-Q Plots for Smoke categories (a) never, (b) smokes now, (c) until current pregnancy, (d) once did, not now

B Additional tables

	pair	diff in means	lwr	upr
0-1	never_smokesNow	-8.75	-11.59	-5.91
0-2	never_untilCurrentPregnancy	0.22	-4.82	5.26
0-3	never_onceDidNotNow	1.77	-3.16	6.70
1-2	smokesNow_untilCurrentPregnancy	8.98	3.89	14.06
1-3	smokesNow_onceDidNotNow	10.52	5.54	15.50
2-3	untilCurrentPregnancy_onceDidNotNow	1.55	-4.94	8.04

Table 7: Tukey-Kramer Confidence Intervals