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: Abdul Muqsit Farooqi

December 17, 2023

Contents

1 Introduction

As the research says smoking causes cancer and other serious problem to the health of a person. Therefore, smoking while pregnant can cause harmful effects on the baby. The defect can lead to premature birth and because of it, babies often have health problems, which can also lead to death. Due to smoke health problems can be lower birth weight, weaker lungs than other babies, and can cause tissue damage in the brain.

The dataset involves the data of mothers and their newborn babies. The instructors of Introductory Case Studies at TU-Dortmund provided a sample of 1236 and 2 independent variables. The independent variables include infant survival, birth weight, date of birth, sex, mother's ethnicity, age, education level, height, weight, and smoking status. The primary objective of this project is to perform a descriptive analysis of sample data and to understand the relationship between maternal smoking and a baby's weight, and whether different conditions of smoking lead to changes in the weight of different groups of neonates.

A global test is carried out by one-way analysis of variance (ANOVA) where the hypothesis of groups having the baby's birth weights differ between the categories. A pairwise comparison is performed by a two-sample t-test between the resulting birth weights. The problem of multiple testing is addressed to adjust the significance level using the Bonferroni correction and Tukey's Honest Significant Difference (HSD).

In section 2, the dataset is explained briefly the quality of the dataset and the structure of the descriptive analysis is discussed. In section 3, statistical methods are explained that are used in fulfilling the task of comparing multiple distributions. In section 4, graphical plots such as QQ-plot, bar chart, histogram, and different tests are used to interpret the results. Lastly, all the results are summarized in section 5.

2 Problem Statement

The dataset used in this project is taken from the Stat Lab. The dataset contains the data on the baby's birth weight in ounces and the mother's smoking status. The level of smoking status of the mother is 0, 1, 2, 3, and 9 where 0 = never smoked, 1 = smokes now, 2 = until current pregnancy, 3 = once did, not now, and 9 = unknown where the status of unknown (9) is removed as it is unknown to us. The dataset consists of 1236 observations and 2 independent variables. The baby's weight change on the level of a mother's smoking is observed. The variable name, type, and description of the observed variables are shown in Table 1.

Table 1: Variable types and their description.

Variable Name	Variable Type	Description
bwt	numeric	Birth weight in ounces
smoke	numeric	Smoking status of mother (0 = never, 1 = smokes now, 2 = until current pregnancy and 3 = once did, not now, 9=unknown)

There are missing values in the dataset and to have better quality of the data, the observations of the missing values are dropped from the dataset. The sample size is big enough that helps with not having biased results but the variation can be seen in the sample data. The problem with a big sample size is that we need to improve the quality of the data by dropping some rows because of missing values and if the data have too much variation, it can also lead to misleading or bad results.

3 Statistical Methods

In this Section, various statistical methods are discussed for the analysis of the data. For the calculation and graphical representation of statistical measures R Software version, 4.2.1 (Core-Team, 2022) is used with the package ggplot2 (Wickham, 2016), dplyr (Wickham et al., 2022), and tseries (Trapletti Hornik, 2022).

3.1 Hypothesis Testing

Hypothesis testing is a statistical inferential method that provides evidence to either accept or reject a belief about some population parameter (Akinkunmi, 2019, p. 141). A belief, also known as a hypothesis, consists of statements in a research study that are compared and contrasted based on sample data collected from a population. These hypothesis statements are always exclusive to each other, and the outcomes of testing such exclusive statements help in making decisions about whether the sample supports a hypothesis regarding the characteristics of the population.

3.1.1 Null Hypothesis and Alternate Hypothesis

The null hypothesis is initially assumed to be true unless it is rejected by strong evidence. The null hypothesis is often represented by Ho. The alternate hypothesis often denoted by $H1$ is the alternative to the null hypothesis. If the sample data shows the evidence for the null hypothesis, then the test accepts the null hypothesis and if the sample data shows the evidence for the alternative hypothesis then the test rejects the null hypothesis.

3.1.2 Significance Level

The significance level is denoted by α. It defines a threshold value that determines if the test is statistically significant or not by the evidence provided by any test. Depending on the type of research study being conducted, the measure is often set in advance. The most common values assigned to α are 0.01, 0.05, 0.1. A significance level of 0.05 indicates that there is a chance of having 5 cases might result in rejecting the null hypothesis out of 100.

3.1.3 P-value

The probability value is also referred to as the p-value. It is the probability of obtaining the smallest value of α for which the null Hypothesis (Ho) can be rejected. P- value ranges between 0 and 1. The smaller the value of p, the greater the chances of the null hypothesis being rejected. A p-value $\leq \alpha$ indicates strong evidence that a null hypothesis can be rejected. Alternatively, if the p-value $> \alpha$ then the null hypothesis is not rejected.

3.1.4 Type I and Type II Errors

Type-I and type-II errors occur when testing a hypothesis. When there is enough evidence for not rejecting the null hypothesis but is rejected then Type-I error is also referred to as a false positive error. The probability of rejecting the null hypothesis when P(rejecting $Ho \mid Ho$ is true) $= \alpha$. Type-II error occurs when there is enough evidence for rejecting the null hypothesis but is not rejected. This is often denoted as β and the β represents the probability of failing to reject a null hypothesis when the null hypothesis is false. i.e. P(fail to reject $Ho \mid Ho$ is false)$= \beta$. Table 2 shows the possible outcomes that can occur during hypothesis testing (Akinkunmi, 2019, p. 142).

Table 2: Statistical errors.

	Ho is false	Ho is true
Reject Ho	True	Type I error
Fail to reject Ho	Type II error	True

3.2 One-Way ANOVA Method

One-way ANOVA (Analysis of Variance) is used to check the statistically significant difference between the means of three or more independent groups. One-Way ANOVA requires a categorical independent variable to have three or more levels and a continuous dependent variable. Based on the levels, the dataset is then divided into groups and are compared using a hypothesis test. The null and alternate hypotheses for one-way ANOVA can be expressed as follows:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_i$$

$$H_1 : \text{At least one of } \mu_i \text{ is different}$$

where $i = 1, 2, \ldots, n$ groups and μ_i represents the mean of the ith group.

The means of all the groups are equal if the null hypothesis is true. If the null hypothesis is false, at least one mean of the group is not equal to the other. The one-way ANOVA test does not suggest which mean of a group is different in the case when the null hypothesis is rejected.

$$F - value = \frac{Groupvariation}{Within\, group\, variation}$$

This test statistic is F-distributed with two degrees of freedom. A small F-value implies that the group means have low variability relative to the variability within each group and having a large F-value implies a high variability of group means in relation to within-group variability (Herzog et al., 2019, p. 70). A hypothesis is then accepted or rejected using the p-value when testing for the specified α level. The above expression is represented formally as follows:

$$F_{stat} = \frac{\frac{1}{k-1}\sum_{j=1}^{k} n_j(\bar{y_j} - \bar{y})^2}{\frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y_j})^2}$$

where k represents the number of groups, nj represents the number of observations within group j, Mj is the mean of group j, MG is the mean of all the groups and xij is the ith observation of a group j.

One-way ANOVA is a parametric test that assumes few characteristics about the data. The primary assumptions of a one-way ANOVA test are that it requires the data to be independent and identically distributed (IID) and the other assumption of equal population variance.

3.3 Quantile-Quantile Plots (QQ-Plots)

The quantile-quantile plot also known as QQ-Plot is a graphical tool that can be used to determine whether a collection of data is likely to have come from a theoretical distribution. QQ-plots provide a summary of whether the two variables are coming from a common distribution. If we want to check if the variable came from a normal distribution, a QQ-plot can be used to verify this assumption.

A Q-Q plot is a scatter plot created by plotting theoretical quantiles against the sample quantile. If both came from the same distribution, then the observation would be that the data points are forming a relatively straight line (Heiberger Burt, 2015, p. 152.).

3.4 Multiple Testing Problem

A single two-sample t-test that compares two groups with α of 0.05, implies that the test of making a type-I error of 5%. With each additional test performance, the likelihood of getting such a Type-I error however increases. In other words, if the level of significance α is 0.05 then the likelihood of not getting Type-I error for one single test is 0.95. When carrying out two independent tests, the likelihood of not getting the type-I error is given by:

$$(1 - \alpha).(1 - \alpha) = (1 - \alpha)^2 \approx 0.902.$$

This problem is referred to as multiple-testing problem. The probability of incorrectly rejecting the null hypothesis is given by:

$$1 - (1 - \alpha)^m, \text{ (1)}$$

where m represents the number of tests performed (Herzog et al., 2019, p. 64).

3.4.1 Bonferroni Corrections and Adjustment of Significance Level

Bonferroni correction method is applied to avoid the problem of multiple testing. The Bonferroni correction is an approach for reducing the type-I error. To set the same α-value across all the m independent tests, equation 1 below can be set to 0.05 and is solved for α.

$$\alpha = 1 - (0.05)^{\frac{1}{m}} \approx \tfrac{0.05}{m} \text{ (2)}$$

this would result in the following:

$$p < \frac{0.05}{m} \text{ (3)}$$

A null hypothesis for all the m tests can then either be accepted or rejected based on (3).

3.5 Levene's Test

Levene's test is used to check if the underlying random variables have symmetric distributions that deviate from normality.

Let n_j be the size of sample j for $j = 1, 2, \ldots, k$ and for $i = 1, 2, \ldots, n_j$ let x_{ij} be the i-th observation for the j-th sample. The mean of the j-th sample is denoted by $\bar{x}_j$ and computes absolute deviations.

$$d_i j = |x_i j - \bar{x_j}|$$

Let $n = \sum_{j=1}^{k} n_j$ denote the mean of all the absolute deviations by $\bar{d}$, and the variances and sample means of the absolute deviations are denoted by $s^2_{d_j}$ and $\bar{d}_j$, respectively. Then the test statistic is:

$$F^* = \frac{\frac{\sum_{j=1}^{k} n_j(\bar{d_j} - \bar{d})^2}{k-1}}{\frac{\sum_{j=1}^{k}(n_j-1)s_d j^2}{n-k}}$$

where $F^* \sim F(k-1, n-k)$, and p-value $= P(F \geq F^*)$ (represents the number of tests performed (Levene, H., 1960, p. 278))

3.6 Shapiro-Wilk Test

The Shapiro-Wilk goodness-of-fit test is used to determine if a random sample, X_i for $i = 1, 2, \ldots, n$, drawn from a normal Gaussian probability distribution with true mean and variance, μ_i and σ^2, respectively. That is $X \sim N(\mu_i, \sigma^2)$.

Thus, we test the following hypothesis:

H_0 : The random sample was drawn from a normal population, $N(\mu_i, \sigma^2)$

Vs

H_a : The random sample does not follow $N(\mu_i, \sigma^2)$

The Shapiro-Wilk Test statistics is used to test the hypothesis which is given by:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})}$$

where $x_{(i)}$ are the ordered sample values and a_i are constants that are generated by the expression,

$$(a_1, a_2, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} m)^{1/2}}$$

with

$$(m = m_1, m_2, ..., m_n)^T$$

being the expected values of the ordered statistics that are independent and identically distributed random variables that follow the standard normal, $N(0, 1)$ and V is the covariance matrix of the order statistics. (Shapiro, S. S. and M. B. Wilk (1965)).

3.7 Tukey's Test

For any distinct pair $j, k = 1, 2, \ldots, p$ where $\bar{y_j} \geq \bar{y_k}$, the formula for all possible $100(1-\alpha)\%$ simultaneous Tukey confidence intervals for the differences $\mu_j - \mu_k$ is defined as follows:

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

where s is the standard error. The critical value

$$q_\alpha > 0 \text{ is such that } P(q \geq q_\alpha) = \alpha \text{ with } q \sim q(p, n-p).$$

Alternatively, the corresponding test statistic:

$$q*_{ij} = \frac{\sqrt{2}(\bar{y_i} - \bar{y_j})}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

can be used to calculate p-value = P(q $\geq q_{*}jk$)

For Tukey's procedure, by design, the individual and joint significance levels are the same, so the p-value for each pairwise difference is compared against α (John W. Tukey, 2003, p. 278).

3.8 Pairwise t Test

A paired t-test is a statistical test also known as a dependent that compares the standard deviations and means of two related groups to determine if there is a significant difference between two groups. A significant difference occurs when the differences between groups are unlikely to be due to sampling error or chance (Dunnett, C. W. 1980). The null and alternative hypothesis for all the pairs is represented as:

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

The alternate hypothesis is defined as:

$$\text{H1} : \mu_i \neq \mu_j \text{ with } i \neq j$$

4 Statistical Analysis

In this section, all statistical methods are used to make a meaningful analysis of the dataset. The categories of the variable are referred to as levels to make better sense of a statement in the further sections of this report.

4.1 Descriptive Analysis of the Variables

In this section, we are performing the descriptive analysis of the two variables which are weight and smoke. From the dataset, we have the details of how the child's weight is varying with the level of smoking. Table 3 below shows the data summary of the variables of change in weight by the level of smoking of a mother.

Table 3: Data summary

Variable name	**Min**	**1st Qu**	**Median**	**Mean**	**3rd Qu**	**Max**	**NA**
Child's weight if the mother never smoked	55	113.75	124.0	122.86	132.25	176	4
Child's weight if the mother smokes now	58	102.00	115.0	114.11	126.00	163	1
Child's weight if the mother smoked until current pregnancy	62	112.00	122.0	123.08	136.50	163	4
Child's weight if the mother has smoked once	65	112.00	124.5	124.63	138.00	170	1

For the above summary, the Min column represents the minimum child's weight in any category is 55. The 1st quartile is a measure of central tendency that indicates the value below which 25% of the data falls indicating the lower range of values. For instance, in the category "child's weight if the mother smokes now" then the first quartile is 126.00. For this data, the median of a child's weight varies across the categories ranging from 115.0 to 124.0 whereas the mean of a child's weight ranges from 114.11 to 124.63. The 3rd quartile is the value below which 75% of the data falls indicating the upper range of values. For example, in the category "child's weight if the mother has smoked once" then the third quartile is 138.00. The Max column where the maximum child's weight in any category ranges from 163 to 176. The NA column of the category "child's weight if the mother smoked until current pregnancy" has 4 missing values.

4.1.1 Frequency Distribution of Variables

Figure 1 given in the appendix represents a histogram of the child's weight according to the mother's smoking habits. It represents the normal distribution for the categories of mothers' smoking habits that are never, current, until pregnancy, and once. The y-axis represents the frequency or count of observations falling within each interval or bin. By the values on the x-axis, we can see that the distribution illustrates most of the values for the mother's smoking habits are between 100-140.

Figure 2 given in the appendix represents a bar chart of the frequency of the smoking habits of women where the y-axis represents the count of the kind of smokers. The x-axis represents the kind of smokers where the kind of smokers are 0 = never, 1 = smokes now, 2 = until current pregnancy and 3 = once did, not now.

4.2 Global Test using One-Way ANOVA Method

As discussed in section 3.2, one-way ANOVA requires an independent variable to have at least three or more levels and a continuous dependent variable. The dataset consists of a category variable that has four different categories and weight as the continuous dependent variable. In this case, the null hypothesis is that there is no significant difference between the mean times of different kinds of smoking, the alternative hypothesis states that there is a difference between the mean times in one or more kinds of smoking.

$$H_0 : \mu_{\text{never}} = \mu_{\text{smokesnow}} = \mu_{\text{untilcurrentpregnancy}} = \mu_{\text{oncedid, notnow}}$$

The alternate hypothesis can be:

$$H_1 : \mu_i \neq \mu_j \text{ where } i \text{ and } j \text{ can be any kind of smoking given above}$$

4.2.1 Validating the Assumptions

In order to validate the assumptions of the ANOVA test, the results from the Levene test and QQ-plots are considered. The first assumption of equal variance is satisfied by the result of the Levene test. The result of the Levene test is given in Table 4.

Table 4: Results of Levene's test

	Df	F-value	P-value
group	3	2.0519	0.1049

From the above table 4, a p-value of 0.1049 is greater than 0.05 So, we do not reject the null hypothesis therefore the variances are the same.

The second assumption of ANOVA for observations being independent and identically distributed is satisfied by removing observations in the dataset that were part of multiple categories. Lastly, the assumption of normality is discussed by using QQ-plots.

Figures 3, 5, and 6 show that most of the observations fall close to the reference line but not exactly on the line unlikely the QQ-plot of smoke is now in Fig 4. The reason can be because of the small size of the dataset. However, it can be clearly interpreted if the dataset had more observations As the sample size increases, the plots generally exhibit greater stability. Given that the data points remain in proximity to the reference line, it can be inferred that the provided dataset adheres to a normal distribution. To support QQ-plot, the Shapiro-Wilk normality test is also used and the Shapiro-Wilk test states that if the p-value is greater than 0.05 then there is insufficient evidence to reject the null hypothesis and it suggests that the data can be reasonably assumed to follow a normal distribution. If the p-value is less than or equal, then it concludes that the data significantly deviates from a normal distribution.

Table 5: Results of Shapiro-Wilk test

	Shapiro-Wilk test	P-value
never_smoke	0.98187	2.959×10^{-6}
now_smoke	0.99707	0.5476
smoke until pregnancy	0.98275	0.2456
smoke once	0.988	0.5079

For never smoking, the p-value is 2.959e-06. Since the p-value is less than 0.05 then there is sufficient evidence to reject the null hypothesis and the "never smoke" group significantly deviates from a normal distribution. While for now smoke, smoke until pregnancy, and smoke once, the p-value is greater than 0.05 therefore there is insufficient evidence to reject the null hypothesis. The data can be assumed to follow a normal distribution.

Once the assumptions are validated, a global test is conducted to examine potential variations in the weights by establishing a significance level of 0.05.

Table 6: Results of ANOVA

	DF	Sum Sq	Mean Sq	F-value	P-Value
category	3	23932	7977	25.72	3.91×10^{-16}
Residuals	1212	375862	310	-	-

In the above table 6, it can be seen that the F-value is 25.72 and the P-value is 3.91×10^{-16} which is less than 0.05. As a result, the null hypothesis is rejected, which states that the mean weight in all categories is equal. This finding implies that there is a difference in the mean weight of at least one category.

4.3 Pairwise Testing

The results derived from the one-way ANOVA revealed a statistically significant disparity among the means of the groups. However, these findings do not provide specific information regarding which particular group exhibits a distinct mean. A pairwise test is performed in order to identify which specific group differs from the rest. The null and alternate hypothesis for all the pairs is represented as follows:

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

Alternate hypothesis is defined as:

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

where i and j = 0, 1, 2, 3 are the different categories in smoking.

Section 4.2.1 addressed the validation of assumptions, and similar assumptions are anticipated for the pairwise t-test, thereby facilitating the execution of the test.

Tukey's test is a method used for comparing all possible pairs of means. Table 7 below presents the results of Tukey's test.

Table 7: Tukey multiple comparisons of means 95% family-wise confidence level

	diff	lwr	upr	p adj
smokes now - never	-8.7530030	-11.593395	-5.912611	0.0000000
until current pregnancy - never	0.2230994	-4.817277	5.263476	0.9994719
once did not now - never	1.7688889	-3.163166	6.700943	0.7927303
until current pregnancy - smokes now	8.9761024	3.889689	14.062516	0.0000366
once did not now - smokes now	10.5218919	5.542798	15.500986	0.0000004
once did not now - until current pregnancy	1.5457895	-4.944892	8.036471	0.9280679

Table 7 provides the results of multiple pairwise comparisons using the Tukey test. Each row represents a comparison between two groups, and the columns are the difference (diff), lower bound (lwr), upper bound (upr), and adjusted p-value (p adj) for each comparison. This case suggests that all the pairs except "until current pregnancy - never", "once did not now - until current pregnancy", and "once did not now - never" show a significant difference in their means and these pairs have adjusted p-value greater 0.05, therefore, the null hypothesis is accepted.

Table 8: Results of pairwise t-test with P value adjustment method: none

	never	smokes now	until current pregnancy
smokes now	5.0×10^{-15}	-	-
until current pregnancy	0.91	6.2×10^{-06}	-
once did - not now	0.36	6.6×10^{-08}	0.54

Table 9: Results of pairwise t-test using Bonferroni Correction

	never	smokes now	until current pregnancy
smokes now	3.0e-14	-	-
until current pregnancy	1	3.7e-05	-
once did - not now	1	3.9e-07	1

As mentioned in section 3.4, when conducting multiple pairwise tests, the probability of making a type-I error also increases. To account for this concern of multiple testing in the analysis, both Bonferroni Corrections and Tukey's Honest Significant Difference (HSD) method are employed.

The findings presented in Tables 8 and 9 depict the outcomes obtained from applying pairwise t-test with p.adjust.method = "none" and Bonferroni Correction tests. Across both Tables 8 and 9, the p-values for most pair comparisons remain statistically significant, suggesting notable differences in their mean values. However, exceptions arise in the pairs of "never - until current pregnancy", "never - once did, not now", and "until current pregnancy - once did, not now", where the p-values remain non-significant even after employing the correction methods. This implies that the means of these pairs are comparable, regardless of the correction applied. It is worth noting that although the p-values have increased as a result of the correction methods, there is insufficient evidence to accept the null hypothesis for the other pairs, namely "never - smokes now", "smokes now - until current pregnancy", and "smokes now - once did, not now". The confidence interval of pairs mentioned in section 3.7 is interpreted by using Fig 7 and it can be seen that 3 intervals of smokes now - never (1-0), until current pregnancy - smokes now (2-1) and once did, not now - smokes now (3-1) where 0 does not exist then there is a significant difference between the means of the compared groups which supports that the null hypothesis is rejected. On the other hand, there is no significant difference between the means of the compared groups which in our case of three intervals until current pregnancy - never (2-0), once did, not now - never (3-0), and once did, not now - until current pregnancy (3-2).

5 Summary

The dataset involves the data of mothers and their newborn babies. The instructors of Introductory Case Studies at TU-Dortmund provided the data of 1236 observations for two explanatory variables. The change in baby weight by the level of smoking of a mother which is 0 = never, 1 = smokes now, 2 = until current pregnancy and 3 = once did, not now is analyzed to find if there are any significant differences in baby's weight between groups.

The initial analysis involved examining the distribution of the child's weight based on the mother's smoking habits using a histogram and bar chart. Additionally, the count of smoking habits was analyzed across different types of smokers. Following this, a global test in the form of one-way ANOVA was conducted to determine if there were significant differences in the mean weights among the groups. For pairwise comparisons, both Tukey's Honest Significant Difference (HSD) test and pairwise t-test were employed. To address the issue of multiple testing, Bonferroni Correction and Tukey's HSD were utilized. The results of the one-way ANOVA indicated that the means of the groups were not similar, leading to the rejection of the null hypothesis. In the pairwise tests, it was observed that all pairs, except for "until current pregnancy - never", "once did not now - until current pregnancy", and "once did not now - never", exhibited significant differences in mean weights. The application of Bonferroni Correction and Tukey's HSD confirmed that the hypothesis remained unchanged from the pairwise tests. Specifically, only the pairs "until current pregnancy - never", "once did not now - until current pregnancy", and "once did not now - never" displayed similar mean weights, while all other pairs exhibited distinct mean weights.

In future research, the methodologies employed in this project can be expanded to encompass a larger dataset, allowing for the inclusion of additional factors that influence a child's development. For instance, the comparison of results could be extended to encompass the mother's diet, environmental factors, health conditions, and mental well-being. Furthermore, this study could be expanded to explore the impact on the baby's overall health and well-being. By considering these additional variables, a more comprehensive understanding of the factors influencing child development can be achieved.

References

[1] Akinkunmi, Mustapha. 2019. *Introduction to Statistics Using R*. Morgan and Claypool Publishers.

[2] Heiberger, Richard M. and Burt, Holland. 2015. *Statistical Analysis and Data Display*. Springer, New York, NY.

[3] Herzog, Michael, Francis, Gregory, and Clarke, Aaron. 2016. *Understanding Statistics and Experimental Design*. Springer International Publishing.

[4] Heumann, Christian, Schomaker, Michael, and Shalabh. 2016. *Introduction to Statistics and Data Analysis*. Springer International Publishing.

[5] ImmobilienScout24. 2019. *Apartment rental offers in Germany*. Data extract from immobilenscout24. https://www.kaggle.com/corrieaar/apartment-rental-offers-in-germany/metadata.

[6] Kassambara, Alboukadel. 2020. *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.4.0.

[7] R Development Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[8] Rasch, Dieter, Verdooren, Rob, and Pilz, Jürgen. 2020. *Applied Statistics: Theory and Problem Solutions with R*. John Wiley and Sons Ltd.

[9] Wickham, Hadley, François, Romain, Henry, Lionel, and Müller, Kirill. 2021. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.6.

[10] Levene, H. 1960. *Robust Tests for the Equality of Variances*. In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, edited by I. Olkin, 278–292. Stanford University Press, Palo Alto, CA.

[11] Hoaglin, D. C. 2003. *John W. Tukey and Data Analysis*. Statistical Science, Vol. 18, No. 3, 311–318.

[12] Shapiro, S. S. and Wilk, M. B. 1965. *An Analysis of Variance Test for Normality (Complete Samples)*. Biometrika, 52, No. 4, 591–611.

[13] Dunnett, C. W. 1980. *Pairwise Multiple Comparison in the Unequal Variance Case*. Journal of the American Statistical Association, 75, No. 372, 796–800.

Appendix

A Additional figures

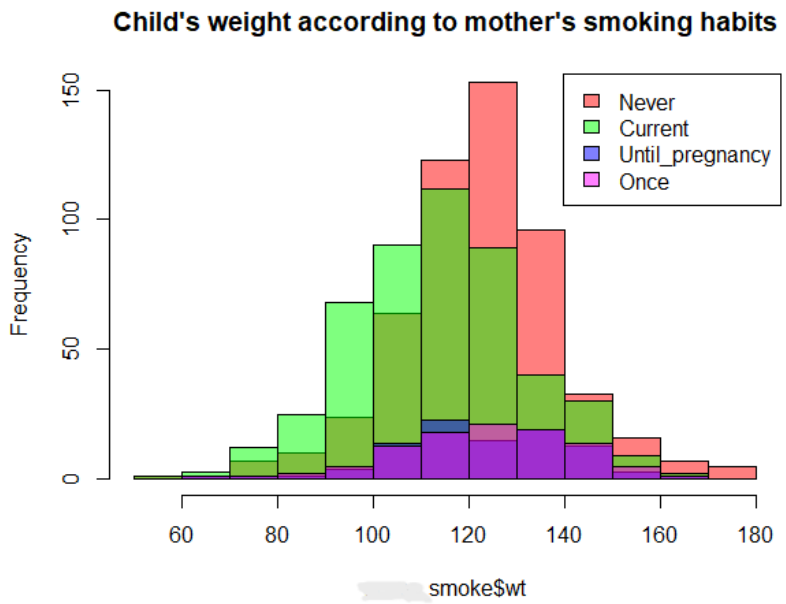

Figure 1: Child's weight according to mother's smoking habits

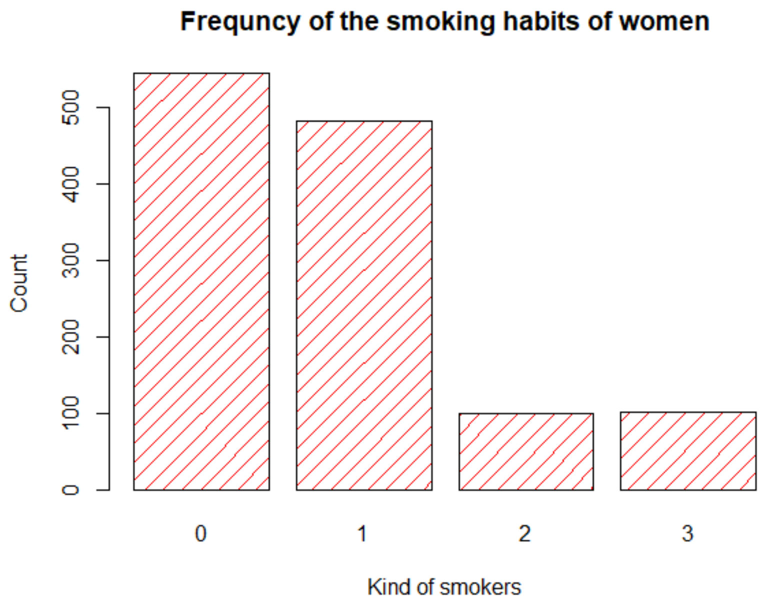

Figure 2: Count of smoking habits of women

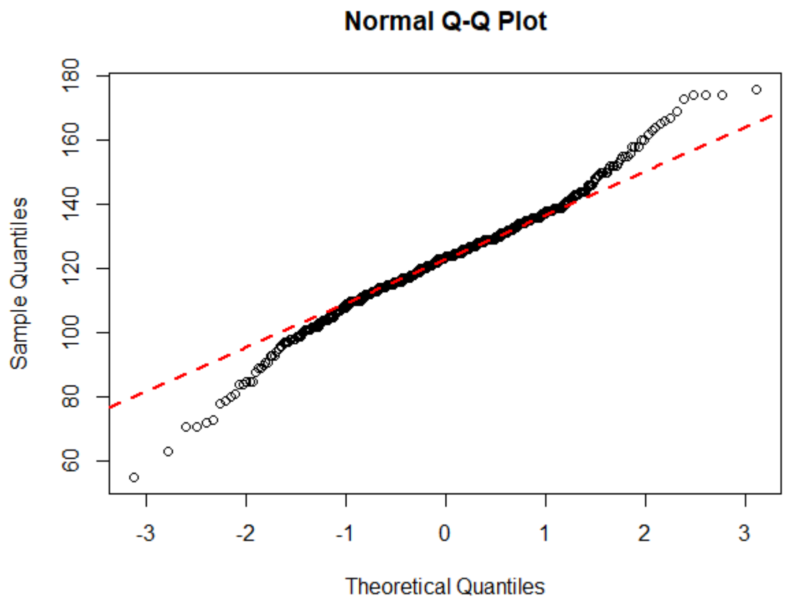

Figure 3: QQ-plot never smoke

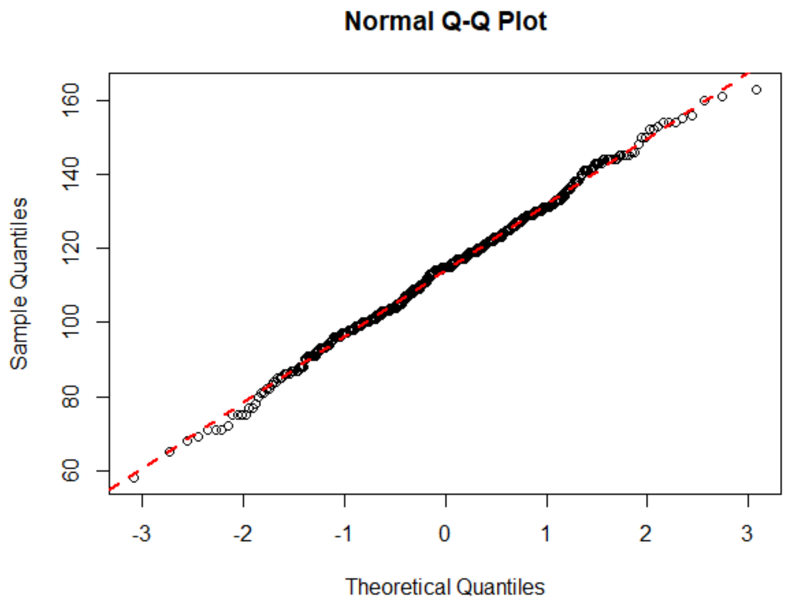

Figure 4: QQ-plot smoke now

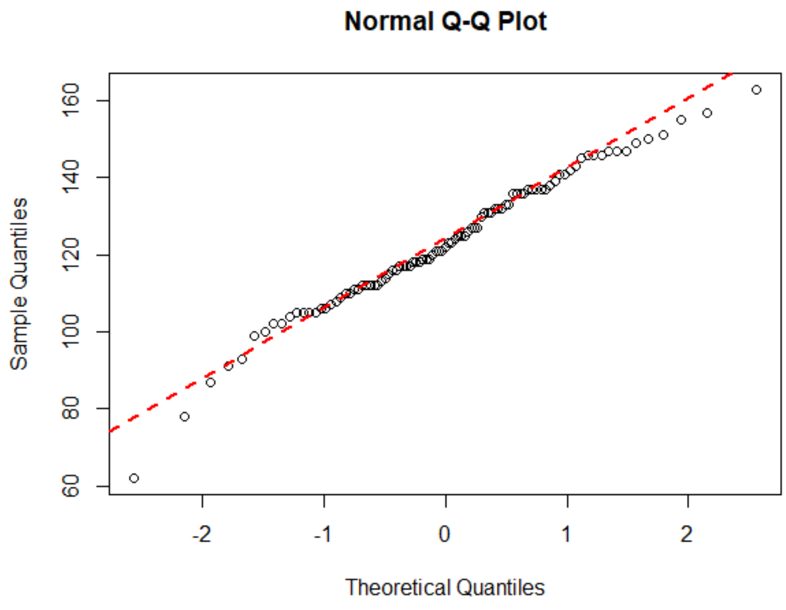

Figure 5: QQ-plot smoke until pregnancy

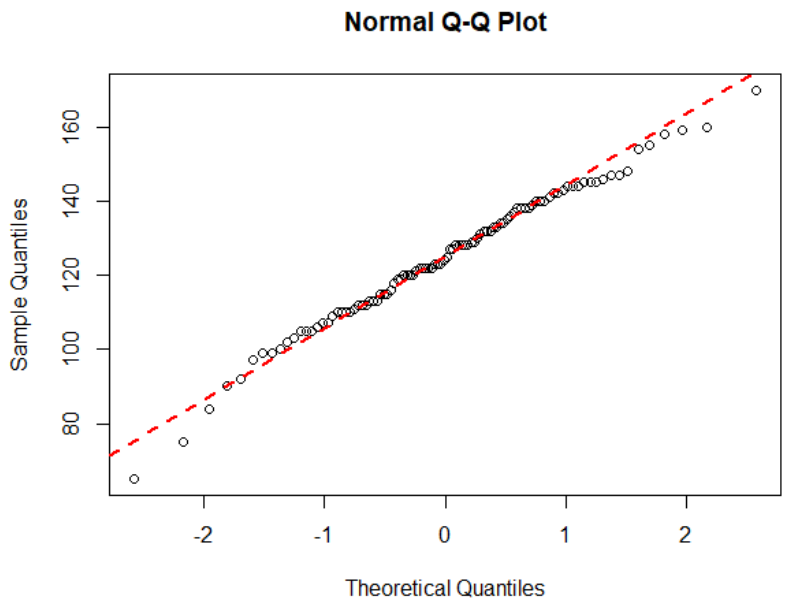

Figure 6: QQ-plot smoke once

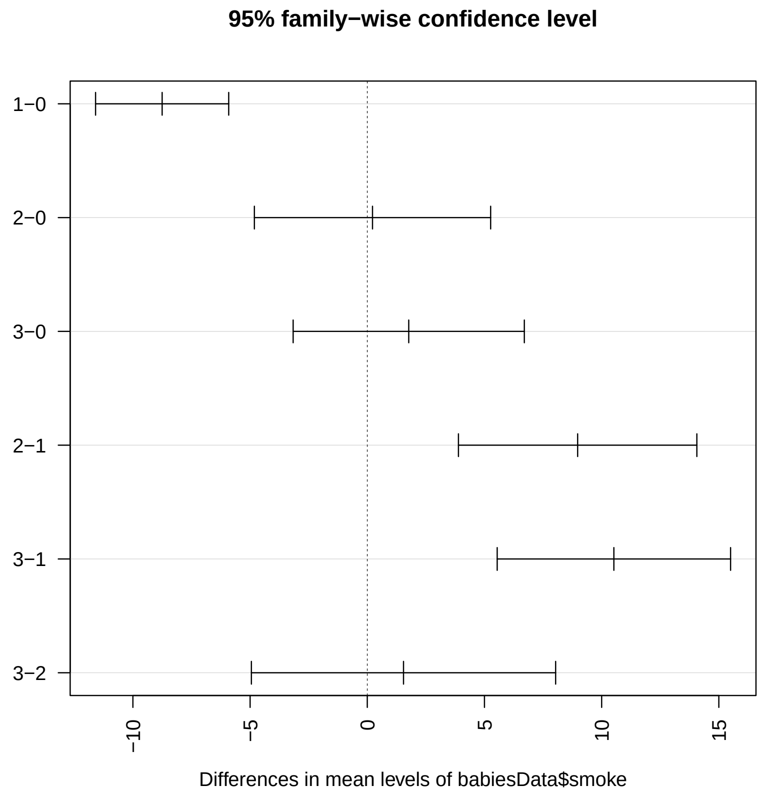

Figure 7: Tukey multiple comparisons of means 95% family-wise confidence level