Filozofická fakulta Univerzity Palackého

Statistical Analysis of Temporal Data in Psycholinguistics

(Bachelor's Thesis)

2024

Tereza Čechová

Filozofická fakulta Univerzity Palackého

Katedra anglistiky a amerikanistiky

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně a uvedla jsem úplný seznam citované a použité literatury.

V Olomouci dne 27. 6. 2024 Tereza Čechová

Acknowledgement

I would like to express my sincere gratitude to my supervisor Mgr. Václav Jonáš Podlipský, Ph.D. for his guidance, help, advice, and support throughout the process of writing this thesis.

Table of Contents

1 Introduction 5

2 Types of Variables 7

3 Statistical Testing and Formal Characteristics of a Hypothesis 9

4 Statistical Tests 11

4.1 Shapiro-Wilk test 11

4.2 T-test 12

4.3 Analysis of Variance (ANOVA) 13

4.4 Mann-Whitney U-test 13

4.5 Kruskal-Wallis Test 14

4.6 Correlation Coefficients 15

4.6.1 Pearson Correlation Coefficient 15

4.6.2 Spearman Rank Correlation Coefficient 16

4.6.3 Kendall Tau Correlation Coefficient 16

5 Language Material 16

6 Hypotheses 18

6.1 The influence of consonant quality in terms of voicing on the duration of the preceding vowel 18

6.2 The influence of consonant quality in terms of voicing on the duration of the preceding vowel considering the type of elicitation 18

6.3 The influence of data elicitation method on the duration of the vowel preceding voiced consonant 19

6.4 The influence of data elicitation method on the duration of vowels preceding voiceless consonants 19

6.5 Differences in consonant duration based on voicing 20

6.6 The relation between vowel and consonant duration 20

6.7 The influence of the speaker's gender 20

6.8 Differences in vowel duration based on consonant voicing - multidimensional analysis 21

7 Results 21

8 Other Methods 36

9 Conclusion 37

References 38
Appendices 40
Appendix 1: R script for the data analysis 40
Appendix 2: R script for plots 47
Annotation 52
Anotace 53

1 Introduction

Statistical analysis has a long history in various linguistic disciplines. It is no different in phonetics. By its very nature, i.e. being a physical quantity, the temporal dimension of speech lends itself to this kind of analysis. Statistical analysis requires thorough preparation, which is typically not included in philological courses. Hence, the aim of this bachelor's thesis is to connect the basics of statistical analysis and phonetics, serving as a basic guide for students interested in conducting their own analyses. A further aim is to demonstrate the application of these methods to a set of data concerning the imitation of English vowel duration variability in speech (Kopecký, 2023).

Linguistics as a discipline allows for the use of a full range of methods, from qualitative to quantitative, and their combinations. Since the early 20th century, when modern linguistics emerged, qualitative methods have been dominant. However, various linguistic disciplines have recently applied quantitative methods more frequently. The first part of this thesis deals with statistical approaches that have become established in linguistic practice, specifically in phonetics. It introduces standard statistical tools such as the Shapiro-Wilk test, t-test, Analysis of Variance (ANOVA), Mann-Whitney U-test, etc. Important aspects of the different approaches are mentioned since choosing the correct method is an essential step in analyzing a given data set. The second part of this thesis focuses on applying selected methods from the theoretical part of this bachelor's thesis on data taken from a master thesis entitled *Imitation of English Coda-Voicing-Induced Vowel Duration Variability by Czech Learners* (Kopecký, 2023). A range of analyses is conducted on this data.

The use of quantitative methods in linguistics is no older than approximately 70 years. (Köhler, 2012) When compared with formal mathematics and logics, which appeared in linguistics around the same time, quantitative methods were established much slower within the field. The first linguist who tried to use quantitative methods in the sense we know it today was George Kingsley Zipf. He started implementing the quantitative methods into qualitative aspects of mathematics and theoretical models. Köhler (2012) says that Zipf's "pioneering work is now considered as the cornerstone of QL [quantitative linguistics]." (Köhler, 2012, p. 13)

Scientific quantitative research in the linguistics field has been getting more attention in the past two decades. Before the 21st century, the majority of linguistics circles were not eager to include statistical methods into their research, arguing that using quantitative evaluation is useless because it is the qualitative aspect of things that is interesting. (Gries, 2013). According to Gries (2013), they did not realize that "quantitative and qualitative methods go hand in hand:

qualitative considerations precede and follow the results of quantitative methods" (p. 4). In empirical studies, counting linguistics as one of them, understanding and choosing correct statistical methods is necessary for conducting a quantitative study. Gries (2013) points out three goals of such studies which are description, explanation, and prediction. The use of statistical methods is crucial in all three steps in order to correctly prepare the data, conduct the analysis, and explain the results as well as predict what future studies might look like.

Further, Gries (2013) demonstrates that simply looking at the data without a deeper statistical analysis is not enough and can lead to incorrect generalizations. Adopting his example, let us illustrate the importance of the use of quantitative methods. He fabricates a small corpus with a distribution of tenses and aspects. The dataset initially seems to corroborate a hypothesis about aspect, which predicts the predominant use of present tenses with imperfective aspects and past tenses with perfective aspects. However, after employing the chi-squared test, a statistical tool suitable for this data, it is revealed that the tense-aspect distribution might merely be caused by chance (to put it more clearly, there is a high probability that the observed co-occurrence is caused by chance). This simple example clearly demonstrates the importance of using statistics in linguistics disciplines as well.

Looking solely at phonetics, Köhler (2012) points out that the use of statistical methods is crucial and phoneticians "could not investigate anything without the measurement of the fundamental quantities like sound pressure, length (duration) and frequency (pitch)" (p. 12). All of these are features which can be represented through numbers. Hence, quantitative analysis is adequate.

Quantitative research is either exploratory in that it looks for prominent patterns in collected data which are used to inform theorizing, or it intends to test theory-driven hypotheses. It is important to follow several steps in order to achieve an accurate analysis. According to Kubát (2016), to test a hypothesis empirically, all the steps can be represented as a cycle. The theory stands at the very beginning. The next step is to formulate a hypothesis derived from the theory. The hypothesis then needs to be formalized (formalizing means converting the hypothesis into the language of numbers, i.e., to express the hypothesis' prediction in terms of measurable quantities so that it can either be confirmed or not by the data collection and analysis). Only after these steps can an experiment be carried out adequately. Hypothesizing after the results are known (referred to as HARKing, Kerr, 1998) is a grave error in research design, potentially leading to biased conclusions. In quantitative linguistics, the experiment usually involves some kind of computation or measurement. The obtained results are then statistically analyzed, determining either the confirmation of the hypothesis or not. One might

think that this is the final step of the research. However, the researcher's work is complete only after interpreting the findings linguistically and connecting them back to the initial theory. The key takeaway is that linguistics stands at both the beginning and the end of research; the experiment is only a tool that allows us to investigate language objectively.

2 Types of Variables

The application of statistical tests is essential to draw meaningful conclusions from the data. Before any of these tests can be carried out, it is crucial to recognize the properties of the variables involved. Addressing these preliminary considerations is important for ensuring the accuracy and reliability of the statistical testing.

This part explores the fundamental process which is involved in classifying the observed phenomena. Classification refers to a systemic process that categorizes the phenomena (e.g., items, characteristics) based on shared attributes. During the process of classification, a value is assigned to each of these phenomena. As a result, a variable is set up. This process allows us to simplify complex data in order to enable a correct understanding and carry out an analysis. There is more than just one way of classifying examined attributes, and each way has an impact on choosing an appropriate statistical test. In other words, the classification is a result of the researcher's decision based on theoretical assumptions. In this thesis, classification is taken from Brzezina (2018) and Köhler (2012).

The first type of variable is called a nominal variable. Nominal variables are classified based on whether or not they possess a certain property. Statements about the observed elements are then evaluated as either true or false within these categories, and there is no hierarchy between the levels of a nominal variable. As a linguistic example, Brzezina takes the category of "[s]peaker's gender [...] because we can assign speakers in the dataset to one of two groups: (1) male speakers and (2) female speakers." (Brzezina, 2018, p.7) For a better understanding, it is possible to formalize the relation as follows:

$$P(A) = P(B) \text{ or } P(A) \neq P(B)$$

where P is the property under consideration, and A and B are two different participants or items (cf. Köhler 2012).

Categorizing observed elements solely on belonging to a group or not is not always sufficient. Hence, in these cases, the type of ordinal variables may be necessary. The similarity

with a nominal variable is that the classification works with distinct categories. However, the categories are not evaluated only as true or false but can also be compared and scaled. Hence, determining whether the "object possesses more, or less, of a given property, or the same amount of it" (Köhler, 2012, p.17) is possible. Brzezina (2018) introduces a linguistic example illustrating this variable, namely, ranked descriptions of a speaker's proficiency level in a foreign language. The labels beginner, pre-intermediate, intermediate, upper-intermediate, advanced, and mastery, as well as the corresponding levels A1, A2, B1, B2, C1, C2, together form the levels of an ordinal variable. Formally:

$$P(A) > P(B),\ P(A) = P(B) \text{ or } P(A) < P(B)$$

where P is the property under consideration, and A and B are two different participants or items.

Likert scale data is an example of a specific type of ordinal data. A Likert scale is commonly used in surveys and questionnaires to measure attitudes, opinions, or perceptions. It allows respondents to express the degree of their agreement or disagreement with a particular statement (strongly agree, agree, neutral, disagree, strongly agree). Jamieson (2004) points out that these "response categories have a rank order, but the intervals between values cannot be presumed equal" (p. 1217)

The third type of variables are scale variables. It involves measuring a particular property and subsequently determining the magnitude of this property. With this approach, the most detailed level of description can be obtained, as it enables the measurement of differences between examined categories and an application of fundamental mathematical operation (addition, subtraction, multiplication, and division). A clear linguistic example in phonetics is the measurement of voice onset time (VOT) because the time between the release of a stop consonant and the onset of voicing can take any value on the scale. Formally:

$$P(A) - P(B) = d$$

where P is the property under consideration, A and B for two different objects, and d stands for the numerical value of the difference.

In some cases, the choice of which variable is used depends solely on the researcher's decision. The operationalization of the observed properties can be chosen arbitrarily. However, this decision impacts not only the selection of adequate statistical methods but also the scope and nature of the conclusions that can be derived from the research. For instance, the variable

of age could be fit for all three types of variables. In terms of the nominal variable, age could be operationalized as a binary category of people who are retired and those who are still in productive age and working. To fit the type of the ordinal variable, age would need to be divided into different levels, such as "children", "young adults", "adults", "the elderly". For the most detailed representation of age, scale variables seem like an ideal choice since they can take on specific values, allowing for precise measurement and analysis.

3 Statistical Testing and Formal Characteristics of a Hypothesis

Statistical testing is a mathematical procedure that allows researchers to interpret hypotheses under consideration. Through hypotheses, one usually attempts to find relationships between selected variables, understanding these relationships as the result of certain general mechanisms. By doing so, it is possible to uncover the underlying principles and links that explain how these variables interact and influence each other. Nonetheless, the hypothesis itself must have specific characteristics in order to be statistically testable.

In empirical research, a hypothesis is not just any prediction considering variables, but it must fulfill specific criteria. Gries (2013) defines the statistically testable hypothesis as a statement that meets the following criteria: "it is a general statement that is concerned with more than just a singular event; it is a statement that at least implicitly has the structure of a conditional sentence *(if ..., then ... or the ..., the ...)* or can be paraphrased as one; it is potentially falsifiable" (p. 11). A falsifiable hypothesis is one that can be proven wrong through empirical evidence, meaning that one must find an event that would contradict the hypothesis. Further, Gries (2013) claims that this approach "implies that the scenario described in the conditional sentence must also be testable. However, these two characteristics are not identical" (p. 11). The difference between these two features comes from the fact that not all the statements that are falsifiable are possible to be tested.

Focusing on specific steps in statistical testing, firstly it is necessary to determine the variables involved, e.g. when studying the relationship between the pronunciation of a specific speech sound and the geographical origin of the speaker, the pronunciation and the speaker's origin are the two observed variables.

As a next step, one needs to formulate the null hypothesis (H0) and the alternative hypothesis (H1). The null hypothesis assumes that there is no relationship between the variables. The alternative hypothesis, on the other hand, suggests that a certain relationship exists between them. This hypothesis should come from a priori theoretical reasoning, which

predicts that the relationship is a consequence of some underlying mechanisms. Continuing with the example of pronunciation and speaker's origin, the null hypothesis states that the geographical origin of the speaker is not associated with the pronunciation of a particular speech sound. The alternative hypothesis states that the regional dialect background and the way speakers pronounce the given speech sounds are related. It should be noted that finding a relationship does not automatically mean proving causality. In other words, the actual outcome of a statistical analysis is essentially just the acceptance or rejection of the null hypothesis.

Based on the chosen statistical test (see below), the validity of the null hypothesis is tested. In other words, it means that we are determining the probability of rejecting the null hypothesis when it is actually true. The probability is assessed through the significance level, which is usually set at 5% (0.05) (assessing the significance level is a crucial step that needs to be done before starting the testing). The significance level determines when the null hypothesis is rejected. Rejecting the null hypothesis with the significance level of 0.05 means that there is a 5% chance that we are wrongfully rejecting it while it is actually true. When the value, usually referred to as a p-value, is higher than the chosen significance level, we do not falsify the null hypothesis. If the p-value is smaller, the null hypothesis is rejected, and the alternative hypothesis is accepted. However, rejecting the null hypothesis does not mean verification of the alternative hypothesis. Moreover, over the past decade, there has been an ongoing discussion about the interpretation and importance of the p-value in empirical research. For instance, Amrhein et al. (2019) represent a strong position against using the p-value in the research, c.f. "Let's be clear about what must stop: we should never conclude there is 'no difference' or 'no association' just because a P value is larger than a threshold such as 0.05 or, equivalently, because a confidence interval includes zero. Neither should we conclude that two studies conflict because one had a statistically significant result and the other did not. These errors waste research efforts and misinform policy decisions." (p. 305-306) A less radical position can be found, for example, in the work of Perezgonzales (2015). The author does not fully reject the approach, but he provides guidelines on how to appropriately apply statistical methods, modifying the original null hypothesis significance testing (NHST) approach and offering alternatives. A critical evaluation of the debate regarding the meaningfulness/meaninglessness of using p-values in empirical research is beyond the scope of this thesis – it is a highly intriguing and complex issue.

4 Statistical Tests

The selection of the appropriate statistical test primarily depends on the nature of the variables which are analyzed. For the analysis of nominal values, which are categorical in nature, researchers are usually concerned with their frequencies. In such cases, the chi-square test or Fisher's exact test are usually the best choices (Gómez, 2013). In terms of scale variables, it is crucial to determine the properties of the distribution of the data because the choice of the appropriate test depends on this property. Specifically, the normality of the distribution is usually tested. This property of the distribution is one of the most important requirements that have to be met for a proper application of the chosen test. According to Razali et. al., "the most commonly used and effective diagnostic tool for checking normality of the data" (p. 21) is the quantile-quantile plot, usually referred to as the Q-Q plot. However, the visualization itself is not always sufficient, and one must rely on more than just the plot. Hence, the normality is further calculated, usually by the Shapiro-Wilk test.

If the condition of the normal distribution of the data is met, it is necessary to use the parametric statistical tests, namely t-test (in case of analyzing two variables), or ANOVA (in case of analyzing multidimensional data). On the other hand, if the condition of the normal distribution of the data is not met, we choose the non-parametric tests, specifically the Mann-Whitney test (in case of analyzing two variables) or Kruskal-Wallis (in case of analyzing multidimensional data). See Table 1.

	Two groups	Multiple groups
Parametric tests	t-test	ANOVA
Non-parametric tests	Mann-Whitney U-test	Kruskal-Wallis test

Table 1: Classification of statistical tests based on their characteristics.

Given the nature of the analyzed data (see below) in the following, section I will briefly characterize only the properties of tests used for scale variables.

4.1 Shapiro-Wilk test

The Shapiro-Wilk test was originally designed to evaluate normality in smaller samples, but since then, it has been extended to be used in larger samples. This test involves fitting ordered

values to a regression line. The Shapiro-Wilk test is particularly important in situations where only a limited number of observations are available, and visualization alone is not sufficient to determine whether the observed values are normally distributed. (Razali et al., 2011) Formally:

$$W = \frac{(\sum_{i=1}^{n} a_i y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where y_i represents the observations (after ordering from smallest to largest), $\bar{y}$ is the sample mean, and a_i are coefficients.

4.2 T-test

The t-test is one of the most common parametric tests which can be used for samples that are independent. It can be applied to two small data sets with independently collected data. As with all statistical tests, the nature of the t-test is to test the validity of the null hypothesis, hence determining if there is any relationship between the observed variables. What is essential for the t-test is the comparison of the means of two groups. However, t-test does not depend solely on the difference between the means. It also takes into account the standard deviations and the sample sizes of the groups. Formally describing the null hypothesis as:

$$H0: \bar{x}_1 = \bar{x}_2$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of two groups. The formula for alternative hypothesis H1 is:

$$H1: \bar{x}_1 \neq \bar{x}_2$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of two groups. Further, the formula for the t-test is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of two groups, s_1 and s_2 are the standard deviations of two groups, and n_1 and n_2 are the sample sizes of two groups.

Required assumptions for this test are as follows: 1. the data have to be normally distributed, 2. the variance of the data needs to be equal or very similar, and 3. the data must have interval scores. (Gómez, 2013).

4.3 Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a statistical test that allows for the comparison of the means of more than two different groups. Similar to the t-test, a fundamental assumption of ANOVA is the normal distribution of the values in given groups. The goal of this statistical test is to determine whether there are statistically significant differences between the means of three or more independent groups.

Formally, the null hypothesis, which assumes no differences among groups under analysis, is:

$$H0: \bar{x}_1 = \bar{x}_2 = \bar{x}_3 = \bar{x}_4 = \bar{x}_5 = \ldots = \bar{x}_n$$

where *n* corresponds to the number of tested groups. The rejection of H0 (and consequently accepting the alternative hypothesis H1) happens in cases when at least one group differs significantly.

The mathematical procedure is more complex as compared to the t-test since it consists of several steps which compute variances and determine whether there are statistically significant differences among the group means (for more details, cf. Gómez, 2013, p. 52-58).

4.4 Mann-Whitney U-test

The Mann-Whitney U-test, also known as the Wilcoxon rank-sum test, belongs to the group of non-parametric tests. Due to its characteristics, it is considered to be an alternative to the t-test, except that the data do not show normal distribution. The computation is based on ranking the data. Specifically, the test uses ranks of the combined data from both groups. (Nachar, 2008) Further, the following formulas are used for computation:

$$U_1 = N_1 N_2 + \frac{N_1(N_1 + 1)}{2} - R$$

$$U_2 = N_1 N_2 - U_1$$

where N_1 is the sample size of the smaller sample, N_2 is the sample size in the other sample, R is the sum of ranks of the smaller sample. For testing statistical significance, we choose the smaller value between U_1 and U_2. (Gómez, 2013)

4.5 Kruskal-Wallis Test

The Kruskal-Wallis test is used for analyzing data of more than two groups. Similar to the Mann-Whitney U-test, it does not require for the data to come from a normal distribution. The null hypothesis of the Kruskal-Wallis test assumes that the measurements in the different groups have the same medians.

$$H0: \tilde{\mu}_1 = \tilde{\mu}2 = \cdots = \tilde{\mu}m$$

where $\tilde{\mu}$ is the median of a specific group, and m is the number of groups.

This test is particularly useful when dealing with non-parametric data or when the assumptions of ANOVA are not met. By ranking the combined data from all groups and comparing the sum of ranks among the groups, the Kruskal-Wallis test determines whether there are statistically significant differences between the groups' medians. If the null hypothesis is rejected, it suggests that at least one group median is different from the others. (Ostertagova et al., 2014)

$$H1: \text{at least for one pair i, j is true that } \tilde{\mu}i \neq \tilde{\mu}j$$

where $\tilde{\mu}$ is the median of a specific group and i and j are two different groups. The complete formula for Kruskal-Wallis test is as follows:

$$H = \left[\frac{12}{n(n+1)}\sum_i \left(\frac{(SR_i)^2}{n_i}\right)\right] - 3(n+1)$$

where n is the total size of the sample, n_i is the specific group's sample size, $(SR_i)^2$ is the total of the ranks related to the specific group. (Hendl, 2004)

4.6 Correlation Coefficients

Correlation is a statistical measure of association between two variables, i.e., the extent to which two variables are related. It indicates how changes in one variable are associated with changes in another. Correlation coefficients, which range from -1 to 1, demonstrate the strength and direction of this relationship. A positive correlation is observed in cases where when one variable increases, the other tends to increase, too. The negative correlation, on the other hand, can be found in cases where when one variable increases, the other tends to decrease. A correlation close to 0 suggests little to no linear relationship between the variables. It is important to realize that correlation does not imply causation. It simply reflects a pattern of association between the variables. (Franzese et al., 2018)

When measuring the strength of associations, the researchers usually choose between three most frequently used methods. In the case of parametric data, the Pearson correlation coefficient is used. When working with non-parametric data, the choice is between the Spearman rank correlation coefficient and the Kendall Tau coefficient. (Gómez, 2013)

4.6.1 Pearson Correlation Coefficient

The Pearson correlation coefficient is a measure of the linear relationship between two continuous variables (interval and rational scales). It quantifies the strength and direction of that relationship. The value of Pearson correlation coefficient, i.e., Pearson's r, ranges from -1 to 1:

$$-1 \leq r \leq 1$$

where r is the Pearson correlation coefficient. If the correlation coefficient equals 1, it indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. (Gómez, 2013) The complete formula for calculating the Pearson correlation coefficient is as follows:

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{\{N\sum x^2 - (\sum x)^2\}\{N\sum y^2 - (\sum y)^2\}}}$$

where r is the Pearson correlation coefficient, N is the number of samples, Σ is the sum, and x and y are two different variables.

4.6.2 Spearman Rank Correlation Coefficient

The Spearman rank correlation coefficient, also known as Spearman's rho, ρ, is a non-parametric measure of the strength and direction of the association between two ranked variables. As opposed to Pearson's correlation, which works with variables measured on an interval scale, Spearman's rank correlation measures the monotonic relationship of variables that are ordinal or data with non-normal distribution. (Gómez, 2013) Formally:

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

where ρ is the Spearman rank correlation coefficient, ΣD^2 is the sum of all squares of the rank differences, and N is the sample size.

4.6.3 Kendall Tau Correlation Coefficient

Similarly to the Spearman rank correlation coefficient, Kendall's tau is a non-parametric statistical test that evaluates the relationship of variables that are ordinal. It is based on the concept of concordant and discordant pairs of observations and is particularly useful for determining the strength and direction of a monotonic relationship between two ranked variables. Compared to Spearman's rho, Kendall's tau is generally more resistant to outliers. (Hendl, 2004) Formally:

$$t_k = \frac{S}{D} = \frac{P - Q}{D}$$

where P is the number of all concordant pairs, Q is the number of all discordant pairs, and D is the maximal number of all possible concordant pairs, resp. discordant pairs (its value is $n(n-1)/2$).

5 Language Material

The data set used in this bachelor's thesis was taken from a master thesis with the title *Imitation of English Coda-Voicing-Induced Vowel Duration Variability by Czech Learners* (Kopecký, 2023). The core of the experiment was to investigate the duration of vowels depending on the voicing of the coda. In this study, various words were analyzed as the participants pronounced them under different conditions. By controlling different aspects of the experiment, such as the elicitation method (shadowing and baseline conditions), the experiment sought to determine whether the voicing has a significant effect on vowel duration. This complex approach allowed

a detailed understanding of phonetic variation and provided insight into the production of speech.

The variable called “session” captures the number of the session in which the data were collected, with two possible values: 1 and 2. In the first session, the participants pronounced words in both, baseline and shadowing condition. In session number 2, only the shadowing elicitation method was used.

The “task” variable categorizes the type of activity participants were engaged in during data collection. The two tasks are shadowing and baseline. Shadowing involves participants repeating words immediately after hearing them, while baseline refers to a condition without such repetition. This distinction helps in understanding how these different speaking conditions affect speech pronunciation characteristics.

The “CodaVoi” variable indicates whether the coda was present or removed. When the coda was removed, the vowel durations were consistent across both voiced and voiceless contexts, suggesting no significant difference in vowel length due to the absence of the coda. In contrast, when the coda was present, the words were analyzed in their original form, considering both English and Czech pronunciations. This differentiation helps in understanding the impact of the coda on vowel duration within different linguistic contexts.

Vowel duration, “vDur”, is a numeric variable representing the duration of the vowel in seconds. The duration range indicates variability in vowel length, which can be analyzed in relation to other variables, such as coda voicing or task.

Coda duration, “cDur”, is another numeric variable that measures the duration of the coda in seconds. This variable helps in understanding the temporal characteristics of the coda (which in this dataset always corresponds with the final consonant, hence, the terms are used interchangeably) in relation to vowel duration.

The “speaker” variable identifies the individual speakers with unique IDs. There are 24 speakers. This variable allows for the examination of individual differences in speech production. The unique subject number assigned to each participant is represented by the variable “sbjNo”.

The “sex” variable indicates the gender of the speaker, with categories woman and man. The predominance of female participants can be considered when analyzing gender-related differences in speech production.

“Age” is a numeric variable representing the age of the participants. The ages range from 19 to 25 years. This relatively narrow age range ensures that age-related variability is minimized, focusing the analysis on other factors.

6 Hypotheses

Based on the data described in Chapter 5 and in accordance with the aim of this bachelor's thesis, which is to explore the possibilities of the application of statistical tests, the following hypotheses are formulated for the data presented in the appendix. Since the goal of this work is to demonstrate the potential application of statistical tests, the simplest approaches are presented first. In these basic analyses, the relationships between only two variables will be examined. Further, more complex analyses that take into consideration three or more variables are introduced. Starting with the simpler methods allows us to lay a foundation for understanding the fundamental relationships in the data and the use of statistical tests. The goal of this step-by-step approach is to present a comprehensive understanding of the possibilities of statistical testing.

In the following part, the analyzed phenomenon is presented first. Next, hypotheses are formulated, and the reasoning behind these choices is described.

6.1 The influence of consonant quality in terms of voicing on the duration of the preceding vowel

H0: The status of the consonant in terms of phonological voicing has no effect on the duration of the preceding vowel.
H1: Vowels that appear before phonologically voiced consonants have a longer duration than those which appear before voiceless consonants.
The differences between voiced and voiceless consonants and their impact on vowel duration are tested regardless of other variables

6.2 The influence of consonant quality in terms of voicing on the duration of the preceding vowel considering the type of elicitation

H0: The status of the consonant in terms of phonological voicing has no effect on the duration of the preceding vowel in a selected group.
H1: Vowels that appear before phonologically voiced consonants have a longer duration than those which appear before voiceless consonants.
The same hypothesis as in case 1. is tested, with the difference being that the data is tested separately for the two distinct categories, i.e., baseline and shadowing. It is based on the

possibility that the method of data elicitation may influence the results. In the baseline condition, where the speaker is not influenced by any other linguistic material, the expected mechanism may or may not appear. Generally, one can assume that in the case of shadowing, there would be a greater influence on the observed mechanism. The reason being that the speakers are directly influenced by the pronunciation patterns of native speakers, potentially leading to more significant differences in vowel duration before voiced and voiceless consonants.

6.3 The influence of data elicitation method on the duration of the vowel preceding voiced consonant

H0: The method of data elicitation has no effect on the duration of vowels preceding phonologically voiced consonants.
H1: Duration of vowels preceding phonologically voiced consonants differ with regard to the character of data elicitation.
This hypothesis tests the differences in vowel duration only before voiced consonants. The impact of the data elicitation method is examined, i.e., the vocal duration is compared both in the baseline group and in the shadowing group. Unlike with analysis 2, the influence of elicitation on the underlying mechanism is not observed - that is, it does not examine the differences in vowel duration based on whether the following consonant is voiced or voiceless. Instead, it focuses on whether the method of elicitation affects the duration of vowels within the same context (in this case, the context refers to the voicing of consonants following the vowel). By isolating the variable of the elicitation method, a more detailed view into whether the conditions of the experiment, such as baseline and shadowing, have a significant effect on vowel length. This is important in order to understand how external factors, e.g. the native speaker's pronunciation patterns in shadowing, can alter the participant's speech patterns compared to a more neutral baseline.

6.4 The influence of data elicitation method on the duration of vowels preceding voiceless consonants

H0: The method of data elicitation has no effect on the duration of vowels preceding phonologically voiceless consonants.
H1: Duration of vowels preceding phonologically voiceless consonants differs with regard to the character of data elicitation.

The same hypothesis as in 3 is tested. However, this one is focused on the duration of vowels before voiceless consonants. The aim is to investigate whether the method of data elicitation influences the vowel duration in this specific context.

6.5 Differences in consonant duration based on voicing

H0: There is no difference in the duration of phonologically voiced and voiceless consonants.
H1: Phonologically voiced consonants are shorter than phonologically voiceless consonants.
This hypothesis is based on the assumption that vowels preceding voiced consonants are typically longer. Hence, the total duration of vowels and consonants should be approximately the same length, c.f.: "The vowel before a voiced coda is longer, and the coda itself is shorter, whereas the vowel preceding a voiceless consonant is shorter and the voiceless coda seems to compensate for this by being slightly longer." (Kopecký, 2024, p.19)

6.6 The relation between vowel and consonant duration

H0: There is no relation between the duration of vowels and consonants.
H1: The longer the vowel, the shorter the consonant.
As in the case of hypothesis 6.5, this hypothesis is based on the same assumption. However, it provides a more detailed examination of this relationship, as it is looked at within the context of a single word form. This method does not test solely the differences. It predicts that there will be a correlation between the duration of vowels and consonants. Specifically, it is expected that the longer the vowel, the shorter the consonant, and vice versa. This analysis should be applied to individual word forms to reduce the mixed effect. This specific approach helps to avoid the mixed effect that might arise from different words and contexts. It provides a clearer picture of the mechanism.

6.7 The influence of the speaker's gender

All of the presented hypotheses can be tested based on gender as well, as all the categories can be divided accordingly. Just as the data can be split into the conditions of baseline and shadowing, it can also be divided by gender.

6.8 Differences in vowel duration based on consonant voicing - multidimensional analysis

H0: There are no differences in vowel duration across the selected groups of data.
H1: There are differences in vowel duration across the selected groups of data.
This hypothesis involves a multidimensional analysis of four specific data groups. Namely, voiced-baseline, voiceless-baseline, voiced-shadowing, voiceless-shadowing. It is a variant of the test presented in 2. The groups combine the variables of the elicitation method and consonant voicing. This analysis should precede the one presented in 2. If the null hypothesis is rejected, then it makes sense to examine the differences as outlined in 2. However, if the null hypothesis is not rejected, it indicates that these factors do not affect the vowel duration, and there is typically no further testing.

7 Results

The results correspond to the order and areas of the observed topics and analyses discussed in the chapter about hypotheses. This is to ensure a structured and coherent presentation of the data.

Ad 6.1 **The influence of consonant quality in terms of voicing on the duration of the preceding vowel**

Table 2 presents values of vowel durations in contexts with voiced and voiceless consonants, providing several statistical measures: mean, median, standard deviation (SD), and Shapiro–Wilk p-values. Firstly, the mean vowel length for voiced consonants is 0.1906 seconds, while for voiceless consonants, it is 0.1447 seconds. This indicates that, on average, vowels are longer when followed by voiced consonants. Secondly, the median values support this observation. The median vowel length is 0.1862 seconds for voiced consonants and 0.1374 seconds for voiceless consonants, confirming that vowels tend to be longer in the presence of voiced consonants. The standard deviation (SD) values give us insights into the variability of vowel lengths. For voiced consonants, the SD is 0.0713, indicating a wider spread of values around the mean. In contrast, the SD for voiceless consonants is 0.0580, suggesting less variability. This higher variability in the voiced context may reflect the influence of additional factors that affect vowel length. Lastly, to test the normal distribution of the data, one can use a Q-Q plot for visualization (Figure 2). In order to label the data as normally distributed, the ordered data

points are plotted in quantiles against the normal distribution. If the data follow the normal distribution, they lie close to a straight 45-degree line. If there are noticeable deviations from the straight line, the data are probably not normally distributed (Das et al., 2016). However, for an exact measurement of the normality, the use of the Shapiro-Wilk test is required. The Shapiro-Wilk p-values for both voiced and voiceless consonants are less than 0.001, indicating that the distribution of vowel lengths deviates from normality in both cases. This proves that the data is not normally distributed. Consequently, the Mann-Whitney U-test is used for a comparison of differences between vowel durations. Since the calculated p-value < 0.001, the null hypothesis is rejected (at the significance level $\alpha = 0.05$), and we can conclude that, for these samples, vowel durations in contexts with voiced and voiceless consonants are statistically significant.

All the information derived from Table 2 can be visualized using a violin plot (Figure 1). Once again, the voiced category shows a symmetrical distribution with a wider spread around the median, indicating higher variability. In contrast with the voiceless category which has a narrower distribution, indicating less variability. The visualization of the data distribution through Q-Q plots can be seen in Figure 2.

	vDur_voiced	vDur_voiceless
Mean	0,1906	0,1447
Median	0,1862	0,1374
SD	0,0713	0,0580
Shapiro–Wilk, p-value	<0.0001	<0.0001

Table 2: Mean, median, and standard deviation of vowel durations for all data. vDur_voiced means the duration of vowels preceding voiced consonants, vDur_voiceless means the duration of vowels preceding voiceless consonants. P-values represent results of Shapiro-Wilk test which tested normal distribution of the data.

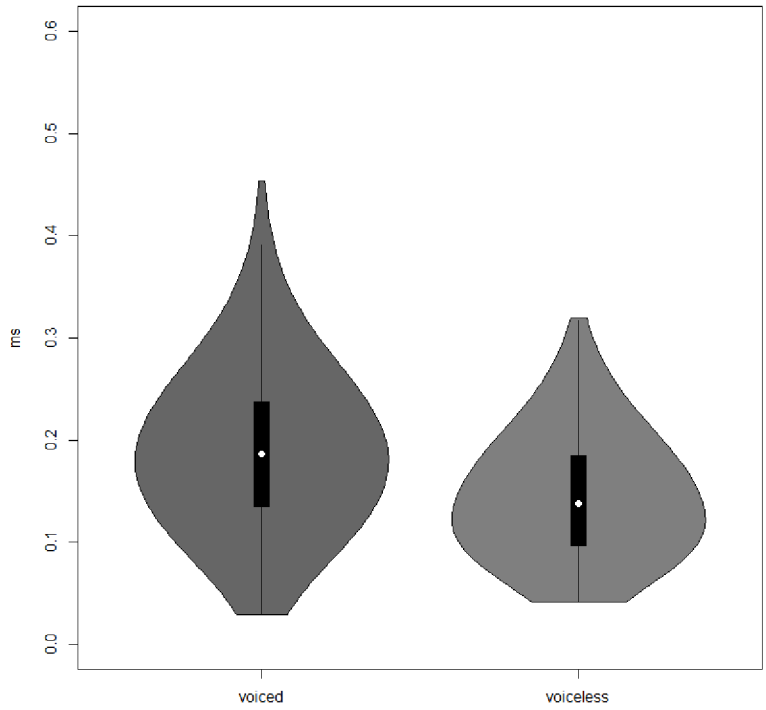

Figure 1: Graphs presenting data from Table 2.

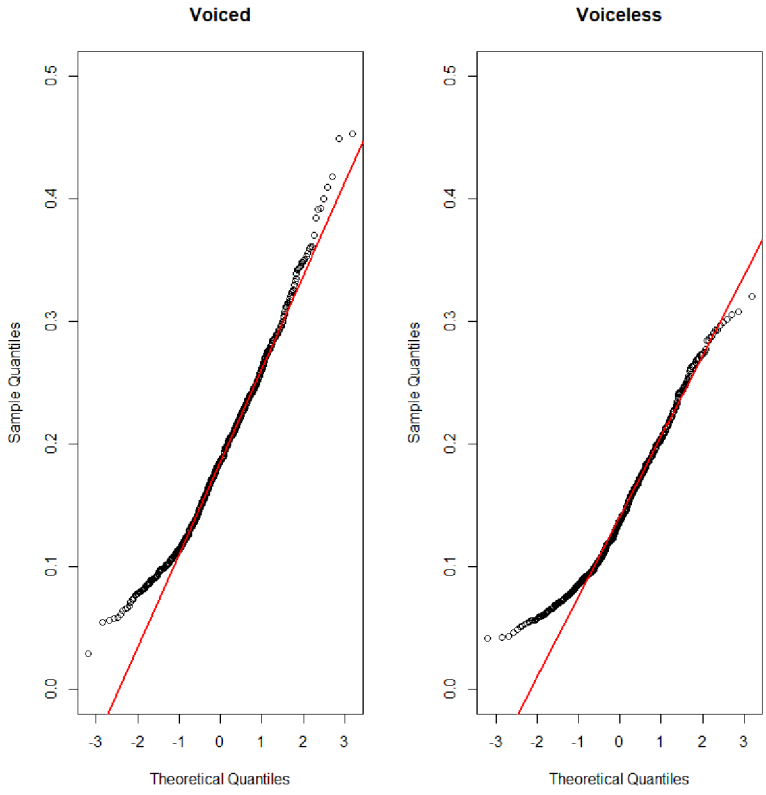

Figure 2: Q-Q plots presenting data from Table 2.

Ad 6.2. **The influence of consonant quality in terms of voicing on the duration of the preceding vowel considering the type of elicitation**

A hypothesis considering the relationship between the vowel duration depending on the context is tested, analogically to the first hypothesis. However, in this case, the data are split into two groups based on the type of elicitation. Meaning that the hypothesis is first tested separately for the baseline data and for the shadowing ones. The aim here is to examine whether the baseline or shadowing conditions influence the validity of the hypothesis.

When examining the first two columns of Table 3, which compare the voiced-baseline and voiceless-baseline, differences can be seen. The mean value for the voiced condition is 0.1750 seconds, which is higher than the mean value for the voiceless condition at 0.1382 seconds. The median values are in accordance with this trend, further supporting the conclusion that, on average, the measurements for the voiced are greater than those for the voiceless. Looking at the standard deviation values, one can observe that it is higher for the voiced (0.0698) than for the voiceless (0.0589). This suggests a slightly greater variability within the data points in the first condition. Lastly, the Shapiro-Wilk test results for both conditions show p-values smaller than 0.001, meaning that both data groups significantly deviate from a normal distribution. As a result, the Mann-Whitney U-test is applied to compare the differences between the voiced and voiceless contexts. Given that the p-value is less than 0.001, the null hypothesis is rejected (at the significance level $\alpha = 0.05$). The conclusion that there is a statistically significant difference in vowel durations between contexts with voiced and voiceless consonants is drawn.

Comparing vowel durations in voiced-shadowing and voiceless-shadowing conditions reveals similar differences as in the case above. Both mean and median values are higher for the voiced-shadowing condition than those for the voiceless. Marginal differences between their standard deviations are again evident, just as in the previous case, with the standard deviation for the voiced condition being slightly larger. The Shapiro-Wilk test p-values are both less than 0.001, indicating that neither group follows a normal distribution. Using the Mann-Whitney U-test to compare the voiced-shadowing and voiceless-shadowing conditions, we find that the calculated p-value is less than 0.001. This leads to the rejection of the null hypothesis (at the significance level $\alpha = 0.05$), concluding that the differences in vowel durations between the voiced and voiceless contexts in shadowing conditions are statistically significant.

Despite the fact that statistically significant differences can be observed in both cases, the difference between voiced-baseline and voiceless-baseline is smaller than the difference between voiced-shadowing and voiceless-shadowing. If we assume that non-native speakers imitate what they hear (what is natural), then shadowing has an impact and the differences in

this category are larger. Specifically, the difference between median values for baseline conditions is 0.0418, whereas for shadowing conditions, it is 0.525. This confirms the assumption that people have a greater tendency to imitate during shadowing. Visualization of the data is presented in Figure 3 and Figure 4.

	v_Dur_voiced - baseline	vDur_voiceless - baseline	vDur_voiced - shadowing	vDur_voiceless - shadowing
Mean	0,1750	0,1382	0,1982	0,1477
Median	0,1643	0,1225	0,1936	0,1411
SD	0,0698	0,0589	0,0708	0,0573
Shapiro–Wilk, p-value	<0.0001	<0.0001	<0.0001	<0.0001

Table 3: Mean, median, and standard deviation of vowel durations based on the type of elicitation. vDur_voiced - baseline means the duration of vowels preceding voiced consonants in baseline conditions, vDur_voiceless - baseline the means duration of vowels preceding voiceless consonants in baseline conditions, vDur_voiced - shadowing means the duration of vowels preceding voiced consonants in shadowing conditions, vDur_voiceless - shadowing means the duration of voiwels preceding voiceless consonants in shadowing conditions. P-values represent results of Shapiro-Wilk test which tested normal distribution of the data.

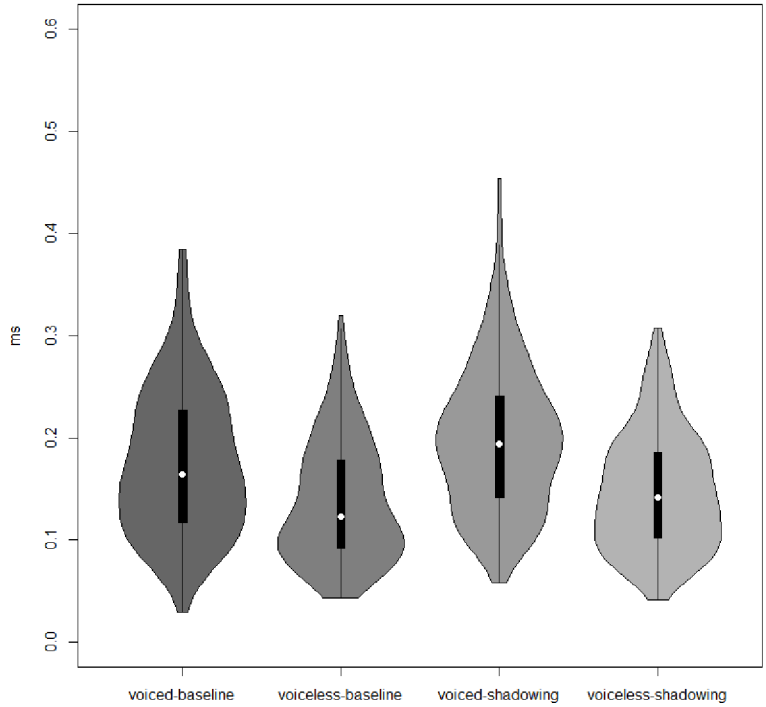

Figure 3: Graphs presenting data from Table 3.

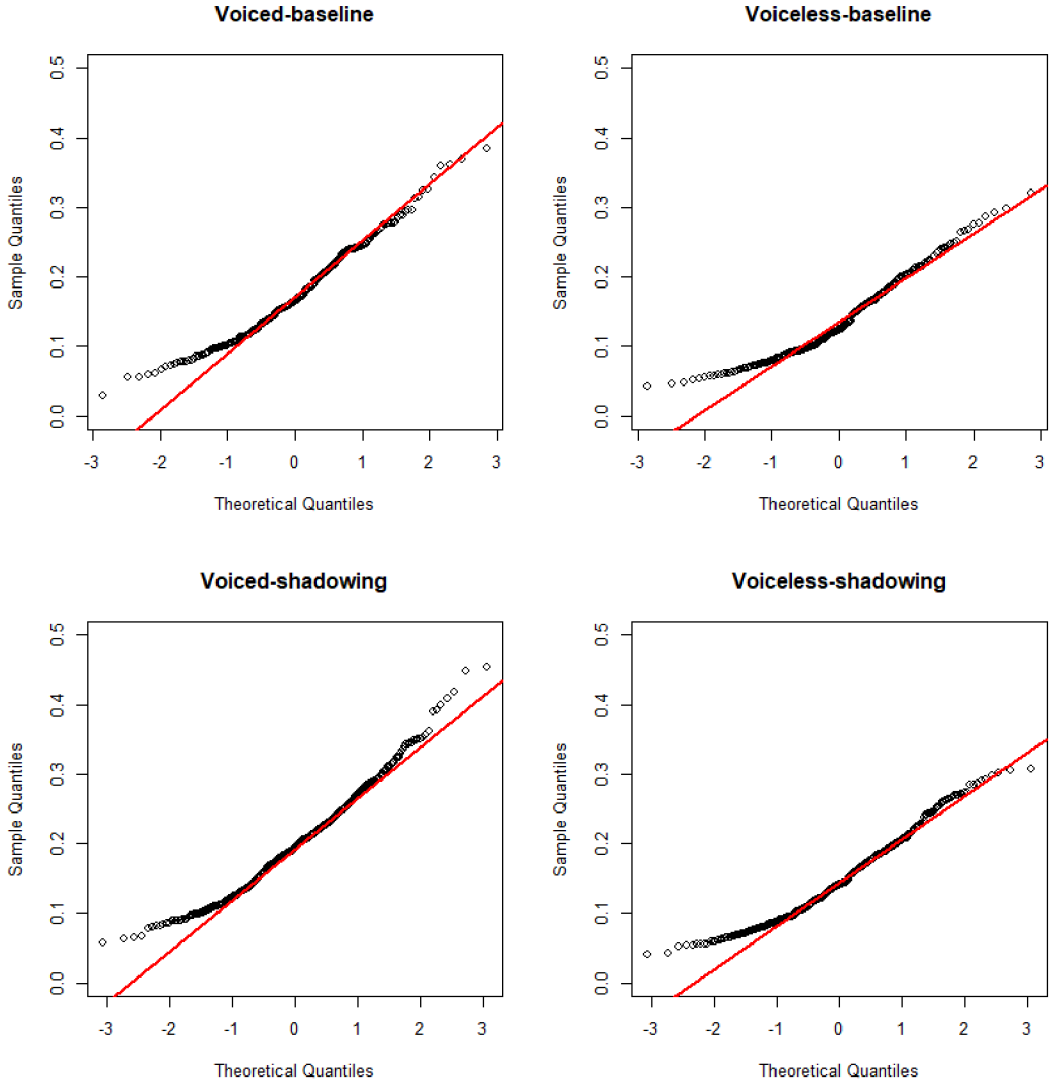

Figure 4: Q-Q plots presenting data from Table 3.

Ad 6.3 **The influence of data elicitation method on the duration of the vowel preceding voiced consonant**

Looking at specific columns from Table 3, the comparison between the "vDur_voiced – baseline" and "vDur_voiced – shadowing" conditions reveals some differences. The voiced - shadowing condition has a higher mean value (0.1982 seconds) compared to the voiced - baseline (0.1750 seconds), indicating an overall increase in the duration under shadowing. The median values also follow this trend, with the shadowing condition at 0.1936 seconds and the baseline at 0.1643 seconds, suggesting that the central value is higher for shadowing. The difference in their standard deviations is minimal, the specific values being 0.0708 for shadowing and 0.0698 for baseline condition. Both groups exhibit statistically significant deviations from normality, as indicated by the Shapiro-Wilk p-values being less than 0.001.

This proves that the data is not normally distributed. The Mann-Whitney U-test is used for comparison (since non-normal distribution of the data). Its p-value < 0.001, i.e., the null hypothesis is rejected (at the significance level $\alpha = 0.05$), and we can conclude that there is a statistically significant difference between vowel duration (in voiced context) in a baseline group, on the one hand, and a group of shadowing, on the other.

Ad 6.4 **The influence of data elicitation method on the duration of vowels preceding voiceless consonants**

Moving onto other columns from Table 3 and comparing solely the "vDur_voiceless - baseline" and "vDur_voiceless - shadowing" conditions reveals distinct differences. The voiceless shadowing condition has a higher mean value (0.1477 seconds) compared to the voiceless baseline (0.1382 seconds), indicating higher vowel durations under shadowing. The median values also reflect this since its value in the shadowing condition is 0.1411 seconds and in the baseline 0.1225 seconds, suggesting a higher central tendency for shadowing. The standard deviation is, as in the previous case, almost the same. However, this time, the shadowing condition is slightly lower (0.0573) than the baseline (0.0589), implying a marginally lesser variability in the data. Both conditions show significant deviations from normality, as indicated by the Shapiro–Wilk p-values being less than 0.001. As in the previous cases, the Mann-Whitney U-test is used for a comparison of differences between vowel durations. The calculated p-value = 0.022, therefore the null hypothesis is rejected (at the significance level $\alpha = 0.05$), and we can conclude that the difference is statistically significant, although the significance level is lower than in previous cases.

Ad 6.5 **Differences in consonant duration based on voicing**

Table 4 exhibits data for an analysis of consonant duration based on voicing, comparing voiced and voiceless consonants. The mean duration of voiced consonants is 0.1229 seconds, while voiceless consonants have a longer mean duration, specifically 0.1535 seconds. Similarly, the median duration of voiced consonants is lower than the one for voiceless consonants, 0.1143 and 0.1475 seconds. Based on the information, a conclusion that, on average, voiceless consonants last longer than voiced consonants can be drawn. The standard deviation of the durations is 0.0487 for voiced consonants and 0.0451 for voiceless consonants, indicating almost the same variability of the data. Lastly, both p-values of the Shapiro-Wilk test are less than 0.001, meaning that the durations do not follow a normal distribution. This proves that the data is not normally distributed. Therefore, the Mann-Whitney U-test is used for a comparison

of differences between consonant durations. The calculated p-value < 0.001 which means that the null hypothesis is rejected (at the significance level $\alpha = 0.05$) and we can conclude that voiced consonants are significantly shorter than voiceless consonants. For the visualization of violin and Q-Q plots, see Figure 5 and Figure 6.

	cDur_voiced	cDur_Voiceless
Mean	0,1229	0,1535
Median	0,1143	0,1475
SD	0,0487	0,0451
Shapiro–Wilk, p-value	<0.0001	<0.0001

Table 4: Mean, median, and standard deviation of coda durations for all data. cDur_voiced means the duration of voiced codas, vDur_voiceless means the duration of voiceless codas. P-values represent results of Shapiro-Wilk test which tested normal distribution of the data.

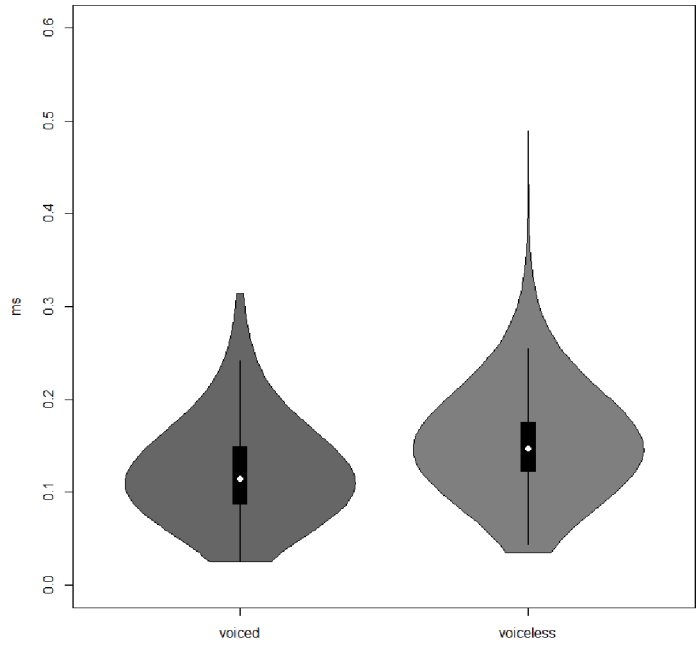

Figure 5: Graphs presenting data from Table 4.

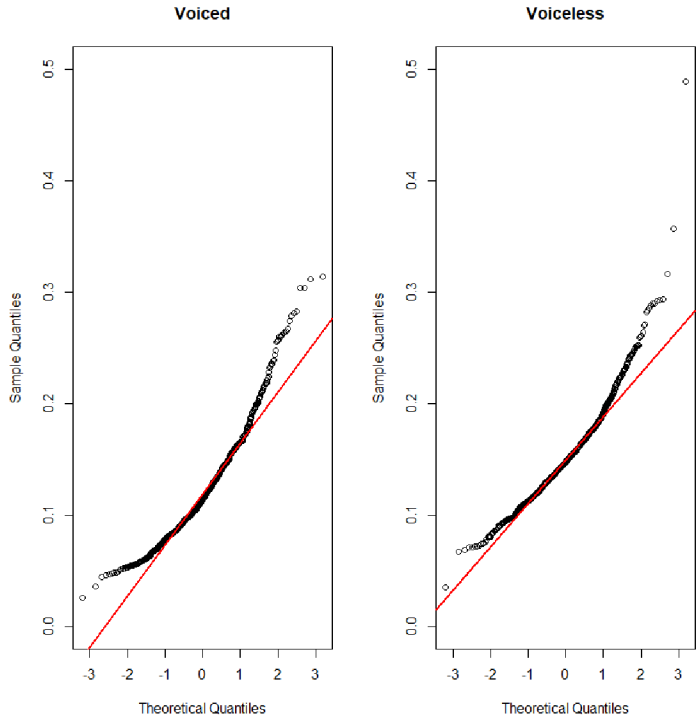

Figure 6: Q-Q plots representing data from Table 4.

Ad 6.6 **The relation between vowel and consonant duration**

According to the hypothesis, the longer the vowel, the shorter the consonant. Thus, it is expected that there should be a negative correlation between durations of vowels and consonants.

First, the correlation was observed for all data. The correlation coefficient measured by Kendall's τ is -0.134. This coefficient indicates a slight inverse relationship between the observed variables. In other words, as one variable increases, the other variable tends to decrease. The weak correlation suggests that while a trend of the variables moving in opposite directions can be observed, the relationship is not strong. However, the significance of the p-value of the correlation indicates that even though it is a weak relationship, it is consistent and reliable across the dataset. Specifically, the p-value for this correlation coefficient is less than 0.001, which is highly significant. Based on this value, one can state that the probability of observing such a correlation by random chance is less than 0.1%. Therefore, the rejection of the null hypothesis of no correlation can be done with a high degree of confidence.

Further, the data are split into two groups, voiced and voiceless. For voiced data, the correlation coefficient τ = -0.103, indicating a weak inverse relationship (just as it is in the case of analyzing all the data). Similar to the overall data, the p-value is less than 0.001, which means this correlation is also statistically significant. This suggests that for voiced data, there is a

significant, however weak, inverse relationship. For the voiceless data, the correlation coefficient τ = -0.028, which is very close to zero. Such a small number indicates almost no relationship. In contrast with the two previous analyses (all data and voiced data), the p-value here is 0.269, which is not statistically significant. This means that there is no evidence of a significant correlation within the voiceless data. Visualizations of correlations are presented in Figure 7 and Figure 8.

	Kendall's τ	P-value
All data	-0,134	<0.001
Voiced	-0,103	<0.001
Voiceless	-0,028	0,269

Table 5: Correlation coefficients between v_Dur and c_Dur measured by Kendall's τ separately for groups of all data, voiced, and voiceless.

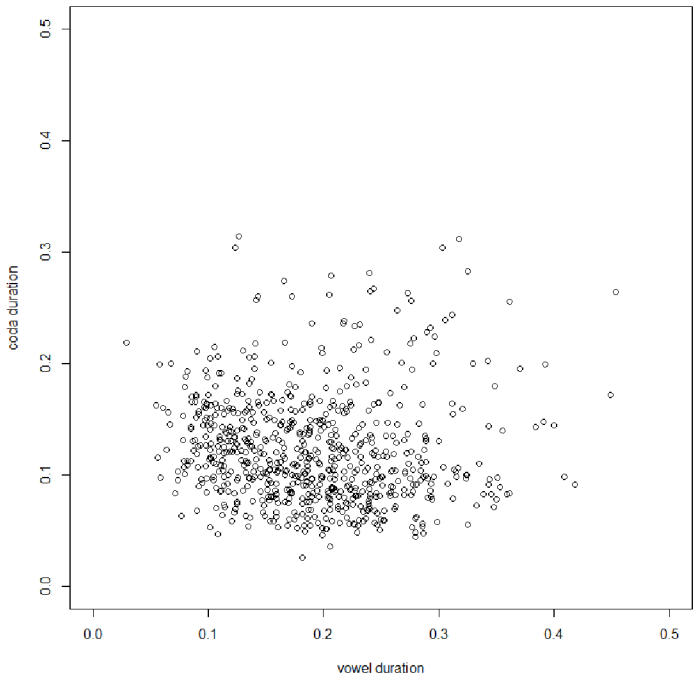

Figure 7: Correlation between v_Dur and c_Dur, all data.

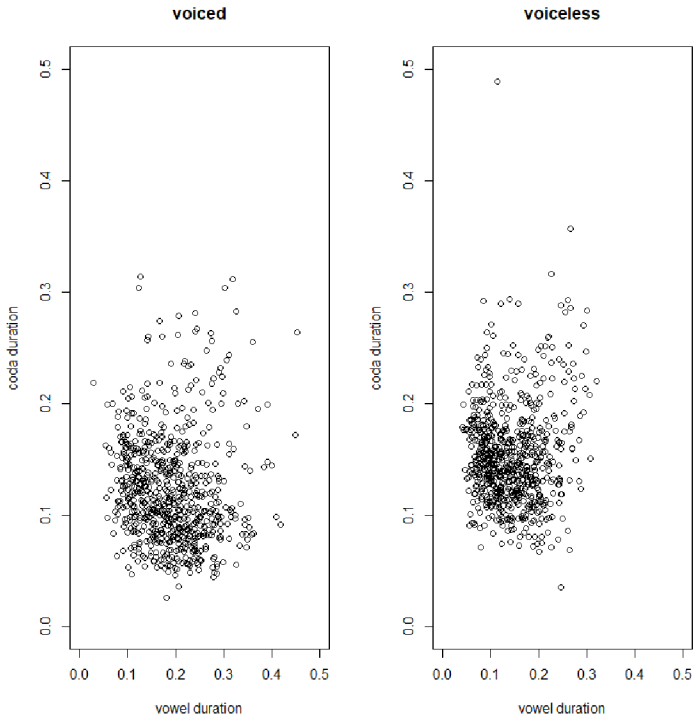

Figure 8: Correlation between v_Dur and c_Dur separately for voiced and voiceless categories.

For a more detailed analysis, the individual words were tested separately (see Table 6). Avoiding the mixed effect was another reason for conducting these individual analyses. When the data were normally distributed, Pearson's correlation coefficient was used to assess relationships. For non-normally distributed data, Kendall's τ method was applied. This approach allows proper capture of the nuances in the data, ensuring that chosen correlation measures are appropriate for the distribution of each dataset.

The values of correlation coefficients and p-values for individual words can be seen in Table 6, with statistically significant correlations highlighted in bold. Specifically, for vowels preceding voiced consonants, there are seven statistically significant correlations and three non-significant ones. In contrast, for vowels preceding voiceless consonants, there is only one statistically significant correlation and nine non-significant ones. The high number of statistically significant correlations for voiced environments suggests that the relationship is strong and consistent. This implies that the following voiced consonants have a meaningful and measurable impact on the duration of vowels. On the contrary, the results for vowels preceding voiceless consonants are quite different. With only one statistically significant correlation and nine non-significant ones, it appears that the voiceless consonants do not have a strong or consistent impact on the duration of preceding vowels. The majority of the relationships for

voiceless vowels are weak or not strong enough to rule out random chance as the deciding factor.

Word	Pearson's ρ	Kendall's τ	P-value	Voicing
bad		-0,174	**0,031**	yes
bat	-0,135		0,258	no
bed		-0,171	**0,038**	yes
bet		-0,119	0,143	no
calf		-0,104	0,205	no
calve		-0,172	**0,038**	yes
cub	-0,306		**<0.001**	yes
cup	-0,129		0,285	no
dock	-0,340		**0,003**	no
dog	-0,313		**<0.001**	yes
gab	-0,174		**0,033**	yes
gap	-0,166		0,165	no
hid		-0,120	0,143	yes
hit	0,036		0,770	no
peck	-0,228		0,054	no
peg	-0,192		0,109	yes
seat	-0,223		0,061	no
seed		-0,051	0,669	yes
tab		-0,254	**0,002**	yes
tap	-0,089		0,273	no

Table 6: Correlation coefficients measured by Pearson's ρ and Kendall's τ for individual words, their p-values, and the status of the consonant in terms of phonological voicing.

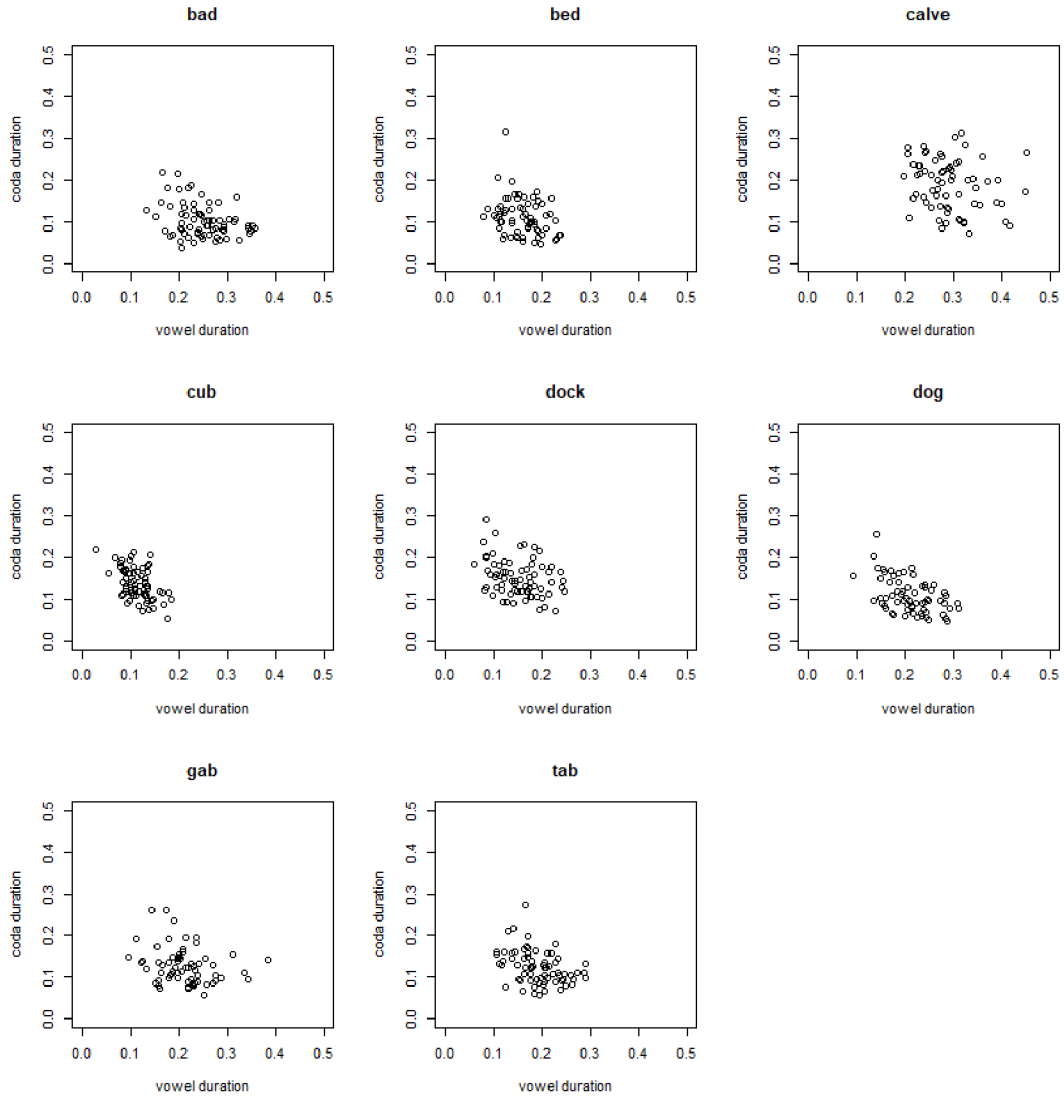

Figure 9: Correlations between v_Dur and c_Dur for individual words (only those with statistically significant correlations).

Ad 6.7 **The influence of the speaker's gender**

It is possible to test all of the mentioned hypotheses based on gender as well. To avoid the mechanical repetition of applying hypotheses to data divided by gender, I will illustrate this by providing only an analysis of the relationship between vowel duration before voiced consonants for both men and women.

The analysis of vowels preceding voiced consonants for men and women reveals differences in their duration. For men, the mean duration of these vowels is 0.1978 seconds, with a median of 0.1915 seconds and a standard deviation of 0.0839. In contrast, values for women are smaller, with a mean duration of 0.1876 seconds, a median of 0.1855 seconds, and a standard deviation of 0.0653. The Shapiro-Wilk p-value is less than 0.001 in both cases,

meaning that data in either of these groups are not normally distributed. The p-value of the Mann-Whitney U-test is 0.2878. Thus, the null hypothesis is not rejected (at the significance level $\alpha = 0.05$). Based on this test, we conclude that there is no statistically significant difference in the duration of vowels in front of voiced consonants between women and men. The very same trend is also visible in case of the duration of vowels before voiceless consonants (p-value = 0.4792). See Figure 10 and Figure 11 for visualization of the data presented in Table 7.

	vDur_Voiced - men	vDur_Voiceless - men	vDur_Voiced - women	vDur_Voiceless - women
Mean	0,1978	0,1483	0,1876	0,1430
Median	0,1915	0,1453	0,1855	0,1341
SD	0,0839	0,0632	0,0653	0,0558
Shapiro–Wilk, p-value	0,001	<0.0001	<0.0001	<0.0001

Table 7: Mean, median, and standard deviations of v_Dur with respect to gender and the status of the consonant in terms of phonological voicing.

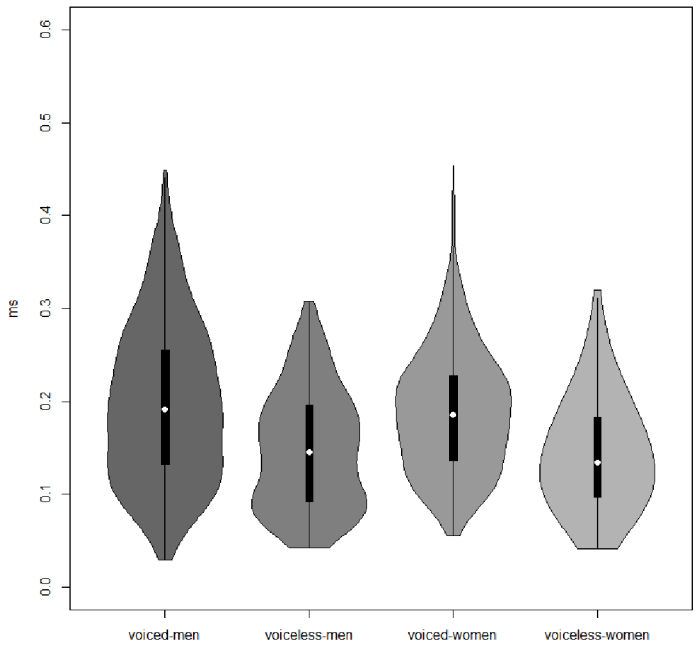

Figure 10: Graphs presenting data from Table 7.

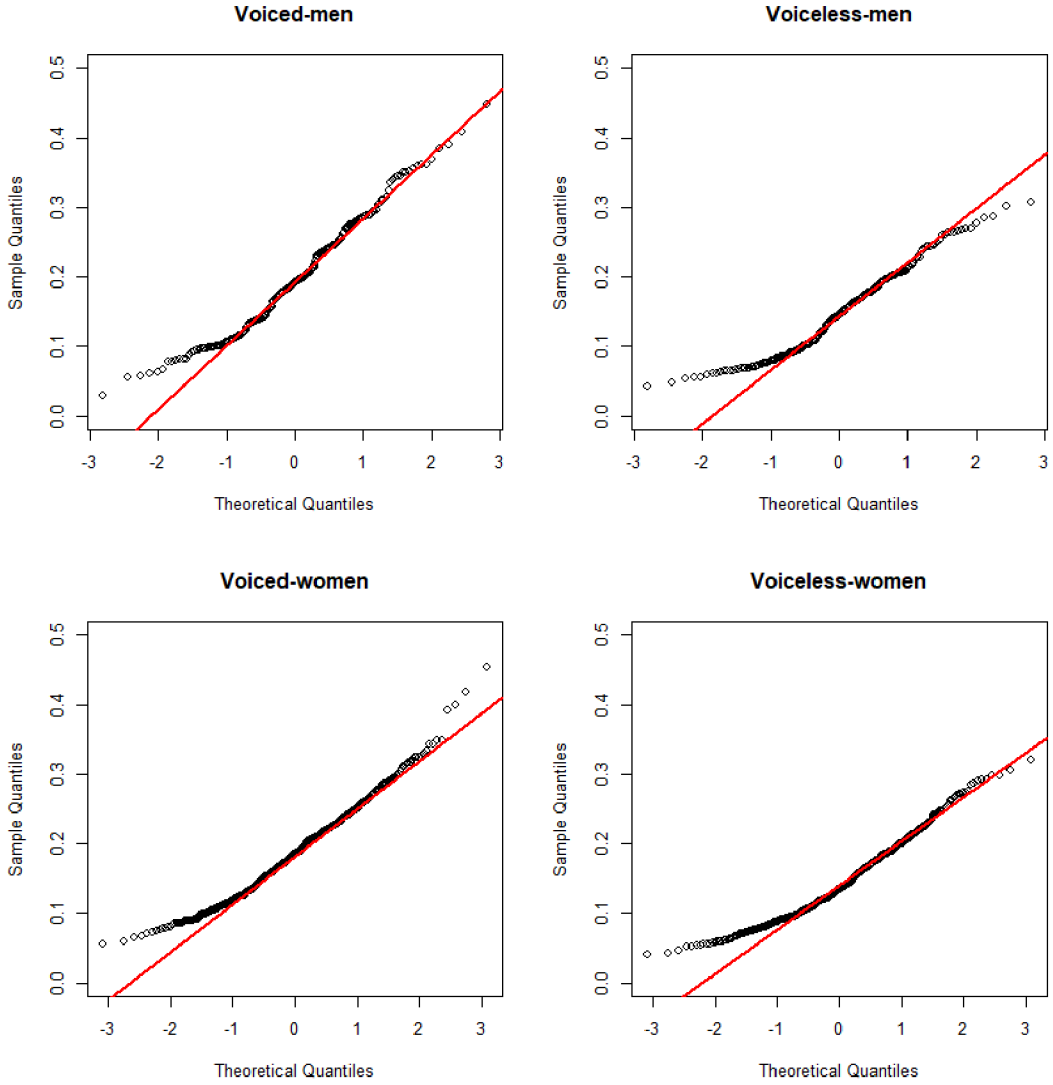

Figure 11: Q-Q plots representing data from Table 7.

Ad 6.8 **Differences in vowel duration based on consonant voicing - multidimensional analysis**

Referring back to Table 3, which presents the statistical measures for four different groups, namely, "v_Dur_voiced – baseline", "v_Dur_voiceless – baseline", "v_Dur_voiced – shadowing", and "v_Dur_voiceless - shadowing". In this section, the results of the multidimensional analysis are presented. In other words, the differences in values within a sample consisting of four different groups are tested. Since the data distribution in each group does not exhibit normal distribution, the Kruskal-Wallis test is used. This non-parametric test gives a p-value less than 0.001. Hence, significant differences in at least one group's mean compared to others occur. It is important to highlight that in the context of analyzing a specific experiment, this test would be used as the initial step in determining statistically significant differences across observed groups.

All the phenomena can be seen also in visualizations in Figure 3 and Figure 4.

8 Other Methods

I would like to emphasize that the methods previously mentioned are far from being the only possible and ideal way of analyzing given material. Like all statistical methods, they have their own limitations and weaknesses. I am well aware that in addition to these methods I have presented, there are other more complex methods commonly used in phonetics. In some cases, these methods are capable of handling the phonetic data more effectively. In this section, I will briefly introduce some of the other methods which are possible to use in these kinds of research.

One of the possible methods of analyzing data that is commonly used in phonetics is called Linear Mixed Models. Matuschek et al. (2017) argue that "[l]inear mixed-effects models have increasingly replaced mixed-model analyses of variance for statistical inference in factorial psycholinguistic experiments." (p. 305) They attribute the shift from ANOVA to linear mixed-effect models (LMMs) to substantial advantages that LMMs have over more traditional methods. One of the most significant benefits is their ability to analyze nested data structures, which are frequently encountered in research with repeated measures or clustered data. Further, LMMs can simultaneously account for both fixed and random effects, which helps in controlling for variability that comes from subject-specific and item-specific differences, leading to more accurate results. They are also better at managing missing data efficiently, ensuring that the integrity of the analysis is maintained even when data is incomplete. Overall, by allowing researchers to explore relationships between variables more precisely than traditional ANOVA, LMMs have become the preferred choice for many researchers. (Matuschek et al., 2017)

Next, Vasishth et al. (2018) introduce another possible method, namely Bayesian data analysis. In their paper, they go into detail about Bayesian linear mixed models in combination with one of the statistical programming languages, R. A step-by-step tutorial of fitting a dataset with information about voice onset times is presented. Apart from this thorough analysis, the advantages of using this method are highlighted as well. As one of the benefits of Bayesian methods, Vashishth et al. (2018) point out that "researchers can (i) flexibly define the underlying process that they believe to have generated the data; (ii) obtain direct information regarding the uncertainty about the parameter that relates the data to the theoretical question being studied." (p. 174) Additionally, the fact that Bayesian models provide flexibility in

defining models and the ability to quantify uncertainty through credible intervals rather than binary significance tests is emphasized.

9 Conclusion

The goal of this bachelor's thesis was to present various statistical methods and their application in phonetic research. Given that this thesis was also intended to serve as a kind of guide for students who are not familiar with statistics and are interested in conducting their analyses. Basic procedures were described and explained. Furthermore, the possibilities of different hypotheses derived from a single dataset were discussed in detail. Eight specific problems were identified, and for each, one or two statistically testable hypotheses were set up. Detailed descriptions of the analyses were provided, along with justifications for the selection of statistical tests.

Initially, the main characteristics of empirical research based on statistical testing were introduced. Further, the significance of understanding the properties of different types of variables involved in statistical testing was discussed. Formal properties that statistically testable hypotheses must be fulfilled were considered. Next, I presented standard statistical tests such as the Shapiro-Wilk test, t-test, Analysis of Variance, Mann-Whitney U-test, etc. The character of each test, its limits, and specific contexts of use were discussed, as well as cases where these tests are appropriate and where not. The application of these methods was demonstrated using data from Kopecký's (2023) master thesis on the imitation of English vowel duration variability by Czech learners. Finally, other possible tests were briefly introduced.

References

Amrhein, V., Greenland, S., McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 567(7748), 305-307.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.

Das, K. R., Imon, A. H. M. R. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics*, 5(1), 5-12.

Franzese, M., Iuliano, A. (2018). Correlation analysis. In *Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics*. Vol. 1. 706-721

Gómez, P. C. (2013). *Statistical methods in language and linguistic research*. Shefield, Equinox Publishing Ltd.

Gries, S. (2013). *Statistics for Linguistics with R: A Practical Introduction*. Berlin, Boston: De Gruyter Mouton.

Hendl, J. (2004). *Přehled statistických metod zpracování dat: analýza a metaanalýza dat*. Praha: Portál.

Jamieson, S. (2004). Likert scales: how to (ab)use them. *Medical Education*, 38: 1217-1218.

Kerr, N. L. (1998). *HARKing:* Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, 2(3), 196-217.

Kopecký, D. (2023). *Imitation of English Coda-Voicing-Induced Vowel Duration Variability by Czech Learners* (master's thesis). Available at: https://theses.cz/id/2bmgxk/ [27.06.2024]

Köhler, R. (2012). *Quantitative Syntax Analysis*. Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110272925

Kubát, M. (2016) *Kvantitativní analýza žánrů*. Ostrava: Ostravská univerzita.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of memory and language*, 94, 305-315.

Nachar, N. (2008). The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. *Tutorials in Quantitative Methods for Psychology*.

Ostertagova, E., Ostertag, O., Kováč, J. (2014). Methodology and Application of the Kruskal-Wallis Test. *Applied Mechanics and Materials*. 611. 115-120.

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in psychology*, 6, 135153.

Razali, N. M., Wah, Y. B. (2011). Power Comparisons of Shapiro-Wilk, Kolmogorov Smirnov, Lilliefors and Anderson-Darling Tests. *J. Stat. Model. Analytics*. 2(1), 21-33.

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of phonetics*, 71, 147-161.

Appendices

Appendix 1: R script for the data analysis

```
# Loading data
# By executing the following command a pop-up window will open
# select the file cechova_data_for_bc_thesis.csv from you PC
data <- read.csv(file=choose.files(), header=TRUE)

# Ad 6.1 The influence of consonant quality in terms of voicing on the duration of the preceding vowel
mean(na.omit(data$all_vdur_voiced))
median(na.omit(data$all_vdur_voiced))
sd(na.omit(data$all_vdur_voiced))
shapiro.test(na.omit(data$all_vdur_voiced))

mean(na.omit(data$all_vdur_voiceless))
median(na.omit(data$all_vdur_voiceless))
sd(na.omit(data$all_vdur_voiceless))
shapiro.test(na.omit(data$all_vdur_voiceless))

wilcox.test(na.omit(data$all_vdur_voiced), na.omit(data$all_vdur_voiceless))

# Ad 6.2. The influence of consonant quality in terms of voicing on the duration of the preceding vowel
# considering the type of elicitation

mean(na.omit(data$baseline_vdur_voiced))
median(na.omit(data$baseline_vdur_voiced))
sd(na.omit(data$baseline_vdur_voiced))
shapiro.test(na.omit(data$baseline_vdur_voiced))

mean(na.omit(data$baseline_vdur_voiceless))
median(na.omit(data$baseline_vdur_voiceless))
```

```
sd(na.omit(data$baseline_vdur_voiceless))
shapiro.test(na.omit(data$baseline_vdur_voiceless))

wilcox.test(na.omit(data$baseline_vdur_voiced), na.omit(data$baseline_vdur_voiceless))

mean(na.omit(data$shadowing_vdur_voiced))
median(na.omit(data$shadowing_vdur_voiced))
sd(na.omit(data$shadowing_vdur_voiced))
shapiro.test(na.omit(data$shadowing_vdur_voiced))

mean(na.omit(data$shadowing_vdur_voiceless))
median(na.omit(data$shadowing_vdur_voiceless))
sd(na.omit(data$shadowing_vdur_voiceless))
shapiro.test(na.omit(data$shadowing_vdur_voiceless))

wilcox.test(na.omit(data$shadowing_vdur_voiced),
na.omit(data$shadowing_vdur_voiceless))

# Ad 6.3 The influence of data elicitation method on the duration of the vowel preceding
voiced consonant

wilcox.test(na.omit(data$baseline_vdur_voiced), na.omit(data$shadowing_vdur_voiced))

# Ad 6.4 The influence of data elicitation method on the duration of vowels preceding
voiceless consonants

wilcox.test(na.omit(data$baseline_vdur_voiceless),
na.omit(data$shadowing_vdur_voiceless))

# Ad 6.5 Differences in consonant duration based on voicing

mean(na.omit(data$cdur_voiced))
median(na.omit(data$cdur_voiced))
sd(na.omit(data$cdur_voiced))
```

```
shapiro.test(na.omit(data$cdur_voiced))

mean(na.omit(data$cdur_voiceless))
median(na.omit(data$cdur_voiceless))
sd(na.omit(data$cdur_voiceless))
shapiro.test(na.omit(data$cdur_voiceless))

wilcox.test(na.omit(data$cdur_voiced), na.omit(data$cdur_voiceless))

# Ad 6.6 The relation between vowel and consonant duration
# all data
shapiro.test(na.omit(data$all_vdur))
shapiro.test(na.omit(data$all_cdur))

cor.test(data$all_vdur, data$all_cdur, method = "kendall")

# voiced
shapiro.test(na.omit(data$voiced_vdur))
shapiro.test(na.omit(data$voiced_cdur))

cor.test(data$voiced_vdur, data$voiced_cdur, method = "pearson")

# voiceless
shapiro.test(na.omit(data$voiceless_vdur))
shapiro.test(na.omit(data$voiceless_vdur))

cor.test(data$voiceless_vdur, data$voiceless_cdur, method = "kendall")

# bad
shapiro.test(na.omit(data$bad_vdur))
shapiro.test(na.omit(data$bad_cdur))

cor.test(data$bad_vdur, data$bad_cdur, method = "kendall")
```

```
# bat
shapiro.test(na.omit(data$bat_vdur))
shapiro.test(na.omit(data$bat_cdur))

cor.test(data$bat_vdur, data$bat_cdur, method = "kendall")

# bed
shapiro.test(na.omit(data$bed_vdur))
shapiro.test(na.omit(data$bed_cdur))

cor.test(data$bed_vdur, data$bed_cdur, method = "kendall")

# bet
shapiro.test(na.omit(data$bet_vdur))
shapiro.test(na.omit(data$bet_cdur))

cor.test(data$bet_vdur, data$bet_cdur, method = "kendall")

# calf
shapiro.test(na.omit(data$calf_vdur))
shapiro.test(na.omit(data$calf_cdur))

cor.test(data$calf_vdur, data$calf_cdur, method = "kendall")

# calve
shapiro.test(na.omit(data$calve_vdur))
shapiro.test(na.omit(data$calve_cdur))

cor.test(data$calve_vdur, data$calve_cdur, method = "kendall")

# cub
shapiro.test(na.omit(data$cub_vdur))
shapiro.test(na.omit(data$cub_cdur))
```

```
cor.test(data$cub_vdur, data$cub_cdur, method = "pearson")

# cup
shapiro.test(na.omit(data$cup_vdur))
shapiro.test(na.omit(data$cup_cdur))

cor.test(data$cup_vdur, data$cup_cdur, method = "pearson")

# dock
shapiro.test(na.omit(data$dock_vdur))
shapiro.test(na.omit(data$dock_cdur))

cor.test(data$dock_vdur, data$dock_cdur, method = "pearson")

# dog
shapiro.test(na.omit(data$dog_vdur))
shapiro.test(na.omit(data$dog_cdur))

cor.test(data$dog_vdur, data$dog_cdur, method = "pearson")

# gab
shapiro.test(na.omit(data$gab_vdur))
shapiro.test(na.omit(data$gab_cdur))

cor.test(data$gab_vdur, data$gab_cdur, method = "pearson")

# gap
shapiro.test(na.omit(data$gap_vdur))
shapiro.test(na.omit(data$gap_cdur))

cor.test(data$gap_vdur, data$gap_cdur, method = "pearson")

# hid
shapiro.test(na.omit(data$hid_vdur))
```

```
shapiro.test(na.omit(data$hid_cdur))

cor.test(data$hid_vdur, data$hid_cdur, method = "kendall")

# hit
shapiro.test(na.omit(data$hit_vdur))
shapiro.test(na.omit(data$hit_cdur))

cor.test(data$hit_vdur, data$hit_cdur, method = "pearson")

# peck
shapiro.test(na.omit(data$peck_vdur))
shapiro.test(na.omit(data$peck_cdur))

cor.test(data$peck_vdur, data$peck_cdur, method = "pearson")

# peg
shapiro.test(na.omit(data$peg_vdur))
shapiro.test(na.omit(data$peg_cdur))

cor.test(data$peg_vdur, data$peg_cdur, method = "pearson")

# seat
shapiro.test(na.omit(data$seat_vdur))
shapiro.test(na.omit(data$seat_cdur))

cor.test(data$seat_vdur, data$seat_cdur, method = "pearson")

# seed
shapiro.test(na.omit(data$seed_vdur))
shapiro.test(na.omit(data$seed_cdur))

cor.test(data$seed_vdur, data$seed_cdur, method = "kendall")
```

```
# tab
shapiro.test(na.omit(data$tab_vdur))
shapiro.test(na.omit(data$tab_cdur))

cor.test(data$tab_vdur, data$tab_cdur, method = "kendall")

# tap
shapiro.test(na.omit(data$tap_vdur))
shapiro.test(na.omit(data$tap_cdur))

cor.test(data$tap_vdur, data$tap_cdur, method = "pearson")

# Ad 6.7 The influence of speakerâ€™s gender

mean(na.omit(data$men_vdur_voiced))
median(na.omit(data$men_vdur_voiced))
sd(na.omit(data$men_vdur_voiced))
shapiro.test(na.omit(data$men_vdur_voiced))

mean(na.omit(data$men_vdur_voiceless))
median(na.omit(data$men_vdur_voiceless))
sd(na.omit(data$men_vdur_voiceless))
shapiro.test(na.omit(data$men_vdur_voiceless))

mean(na.omit(data$women_vdur_voiced))
median(na.omit(data$women_vdur_voiced))
sd(na.omit(data$women_vdur_voiced))
shapiro.test(na.omit(data$women_vdur_voiced))

mean(na.omit(data$women_vdur_voiceless))
median(na.omit(data$women_vdur_voiceless))
sd(na.omit(data$women_vdur_voiceless))
shapiro.test(na.omit(data$women_vdur_voiceless))
```

```
wilcox.test(na.omit(data$men_vdur_voiced), na.omit(data$women_vdur_voiced))

# Ad 6.8 Differences in vowel duration based on consonant voicing - multidimensional
analysis

values = c(data$baseline_vdur_voiced, data$baseline_vdur_voiceless,
          data$shadowing_vdur_voiced, data$shadowing_vdur_voiceless)
group = factor(c(rep(1, length(data$baseline_vdur_voiced)), rep(2,
length(data$baseline_vdur_voiceless)),
                rep(3, length(data$shadowing_vdur_voiced)), rep(4,
length(data$shadowing_vdur_voiceless))))

data <- data.frame(values, group)
kruskal_result <- kruskal.test(values ~ group, data = data)
print(kruskal_result)
```

Appendix 2: R script for plots

```
library(vioplot)
# Loading data
# By executing the following command a pop-up window will open
# select the file cechova_data_for_bc_thesis.csv from you PC
data <- read.csv(file=choose.files(), header=TRUE)

# fig01
vioplot(data$all_vdur_voiced, data$all_vdur_voiceless, names = c("voiced", "voiceless"),
        col = c("grey40", "gray50"), ylim = c(0,0.6))
title(ylab = "s")

# fig02
par(mfrow = c(1, 2))
qqnorm(data$all_vdur_voiced, ylim = c(0,0.5), main = "Voiced")
qqline(data$all_vdur_voiced, col = "red", lwd = 2)
```

```
qqnorm(data$all_vdur_voiceless, ylim = c(0,0.5), main = "Voiceless")
qqline(data$all_vdur_voiceless, col = "red", lwd = 2)
par(mfrow = c(1, 1))

# fig03
vioplot(data$baseline_vdur_voiced, data$baseline_vdur_voiceless,
      data$shadowing_vdur_voiced, data$shadowing_vdur_voiceless,
      names = c("voiced-baseline", "voiceless-baseline", "voiced-shadowing", "voiceless-
shadowing"),
      col = c("grey40", "gray50", "gray60", "gray70"), ylim = c(0,0.6))
title(ylab = "s")

# fig04
par(mfrow = c(2, 2))
qqnorm(data$baseline_vdur_voiced, ylim = c(0,0.5), main = "Voiced-baseline")
qqline(data$baseline_vdur_voiced, col = "red", lwd = 2)
qqnorm(data$baseline_vdur_voiceless, ylim = c(0,0.5), main = "Voiceless-baseline")
qqline(data$baseline_vdur_voiceless, col = "red", lwd = 2)
qqnorm(data$shadowing_vdur_voiced, ylim = c(0,0.5), main = "Voiced-shadowing")
qqline(data$shadowing_vdur_voiced, col = "red", lwd = 2)
qqnorm(data$shadowing_vdur_voiceless, ylim = c(0,0.5), main = "Voiceless-shadowing")
qqline(data$shadowing_vdur_voiceless, col = "red", lwd = 2)
par(mfrow = c(1, 1))

# fig05
vioplot(data$cdur_voiced, data$cdur_voiceless, names = c("voiced", "voiceless"),
      col = c("grey40", "gray50"), ylim = c(0,0.6))
title(ylab = "s")

# fig06
par(mfrow = c(1, 2))
qqnorm(data$cdur_voiced, ylim = c(0,0.5), main = "Voiced")
qqline(data$cdur_voiced, col = "red", lwd = 2)
qqnorm(data$cdur_voiceless, ylim = c(0,0.5), main = "Voiceless")
```

```
qqline(data$cdur_voiceless, col = "red", lwd = 2)
par(mfrow = c(1, 1))

# fig07
plot(data$all_vdur,data$all_cdur, xlab = "vowel duration", ylab = "coda duration",
    xlim = c(0,0.5), ylim = c(0,0.5))

# fig08
par(mfrow = c(1, 2))
plot(data$voiced_vdur,data$voiced_cdur, xlab = "vowel duration", ylab = "coda duration",
    xlim = c(0,0.5), ylim = c(0,0.5), main ="voiced")
plot(data$voiceless_vdur,data$voiceless_cdur, xlab = "vowel duration", ylab = "coda
duration",
    xlim = c(0,0.5), ylim = c(0,0.5), main ="voiceless")
par(mfrow = c(1, 1))

# fig09
par(mfrow = c(3, 3))

plot(data$bad_vdur,data$bad_cdur, xlab = "vowel duration", ylab = "coda duration",
    xlim = c(0,0.5), ylim = c(0,0.5), main ="bad")

plot(data$bed_vdur,data$bed_cdur, xlab = "vowel duration", ylab = "coda duration",
    xlim = c(0,0.5), ylim = c(0,0.5), main ="bed")

plot(data$calve_vdur,data$calve_cdur, xlab = "vowel duration", ylab = "coda duration",
    xlim = c(0,0.5), ylim = c(0,0.5), main ="calve")

plot(data$cub_vdur,data$cub_cdur, xlab = "vowel duration", ylab = "coda duration",
    xlim = c(0,0.5), ylim = c(0,0.5), main ="cub")

plot(data$dock_vdur,data$dock_cdur, xlab = "vowel duration", ylab = "coda duration",
    xlim = c(0,0.5), ylim = c(0,0.5), main ="dock")
```

```
plot(data$dog_vdur,data$dog_cdur, xlab = "vowel duration", ylab = "coda duration",
    xlim = c(0,0.5), ylim = c(0,0.5), main ="dog")

plot(data$gab_vdur,data$gab_cdur, xlab = "vowel duration", ylab = "coda duration",
    xlim = c(0,0.5), ylim = c(0,0.5), main ="gab")

plot(data$tab_vdur,data$tab_cdur, xlab = "vowel duration", ylab = "coda duration",
    xlim = c(0,0.5), ylim = c(0,0.5), main ="tab")

par(mfrow = c(1, 1))

# fig10
vioplot(data$men_vdur_voiced, data$men_vdur_voiceless, data$women_vdur_voiced,
data$women_vdur_voiceless,
      names = c("voiced-men", "voiceless-men", "voiced-women", "voiceless-women"),
      col = c("grey40", "gray50", "gray60", "gray70"), ylim = c(0,0.6))
title(ylab = "s")

# fig11
par(mfrow = c(2, 2))
qqnorm(data$men_vdur_voiced, ylim = c(0,0.5), main = "Voiced-men")
qqline(data$men_vdur_voiced, col = "red", lwd = 2)
qqnorm(data$men_vdur_voiceless, ylim = c(0,0.5), main = "Voiceless-men")
qqline(data$men_vdur_voiceless, col = "red", lwd = 2)
qqnorm(data$women_vdur_voiced, ylim = c(0,0.5), main = "Voiced-women")
qqline(data$women_vdur_voiced, col = "red", lwd = 2)
qqnorm(data$women_vdur_voiceless, ylim = c(0,0.5), main = "Voiceless-women")
qqline(data$women_vdur_voiceless, col = "red", lwd = 2)
par(mfrow = c(1, 1))
```

Annotation

Author: Tereza Čechová

Field of Study: English Philology & Czech Philology

Title: Statistical Analysis of Temporal Data in Psycholinguistics

Type: Bachelor's thesis

Faculty and Department: Faculty of Arts, Department of English and American Studies

Supervisor: Mgr. Václav Jonáš Podlipský, Ph.D.

Number of pages: 39 (without appendices)

Number of characters: 66 035

Keywords: statistical testing, vowel duration, hypothesis, phonetic research, variables

Description: This bachelor's thesis is concerned with the possibilities of statistical testing within the field of phonetics. It dives into the process of analyzing temporal data, describing and demonstrating the necessary steps that these kinds of analyses require. Taking into account that this thesis is meant to be a kind of guide for students interested in the use of statistics in phonetics, it mostly deals with basic analyses. Before focusing on the analysis itself, characteristics of empirical research, types of variables, hypotheses, and statistical tests are described. Further, based on the data taken from a master thesis *Imitation of English Coda-Voicing-Induced Vowels* (Kopecký, 2023), possible hypotheses are discussed, and corresponding analyses are conducted. Finally, a brief introduction of other possible statistical tests was provided.

Anotace

Autor: Tereza Čechová

Studijní obor: Anglická filologie / Česká filologie

Název: Statistická analýza temporálních dat v psycholingvistice

Typ práce: Bakalářská práce

Fakulta a katedra: Filozofická fakulta, Katedra anglistiky a amerikanistiky

Vedoucí práce: Mgr. Václav Jonáš Podlipský, Ph.D.

Počet stran: 39 (bez příloh)

Počet znaků: 66 035

Klíčová slova: statistické testování, délka trvání vokálu, hypotéza, fonetický výzkum, proměnné

Charakteristika: Tato bakalářská práce se zabývá možnostmi statistického testování v oblasti fonetiky. Zaměřuje se na proces analýzy temporálních dat, popisuje a předvádí nezbytné kroky, které tyto druhy analýz vyžadují. Vzhledem k tomu, že práce má sloužit také jako určitý návod pro studenty, kteří se zajímají o využití statistiky ve fonetice, zabývá se převážně základními analýzami. Před samotnou analýzou jsou popsány charakteristiky empirického výzkumu, typy proměnných, hypotézy a statistické testy. Dále jsou na základě dat z diplomové práce *Imitation of English Coda-Voicing-Induced Vowels* (Kopecký, 2023) předloženy možné hypotézy a provedeny odpovídající analýzy. Nakonec je představen stručný popis dalších možných statistických metod.