Testing statistical hypotheses using non-parametric tests

Sandro Schweiss

IUBH International College

Statistics: Inferential Statistics DLBDSSIS01

Tutor Stefan Stöckl

Abstract

The stating and debunking of hypotheses are one of the core foundations of science for hundreds of years. A hypothesis is a proposed explanation for new phenomena, which is either rejected or failed to reject. That is how science could progress over the centuries and continuously evolved. In the scope of this written assignment, the procedure of hypothesis testing is thoroughly examined from a statistical perspective. The following assignment covers the paradigms, procedures, and key parameters of hypothesis testing. The different methodologies to conduct experiments are explained, and selected non-parametric methods are showcased. To see the application of the acquired knowledge on a hands-on example, an experiment is conducted, which covers the whole process of hypothesis testing.

Table of Contents

I. List of Tables

2. Testing statistical hypotheses using non-parametric tests

The topic of the following assignment is the testing of statistical hypotheses using non-parametric tests, which is conducted in the scope of my study course. The testing of statistical hypotheses was my main interest in picking this topic since it is a critical component of statistical interference and is relevant for further research and experiments. Also, non-parametric tests piqued my interest more than other methods because of their unique methodology. The purpose of this assignment is to elaborate on the general framework and interpretation of statistical testing, with a focus on non-parametric tests. At first statistical hypotheses testing and key definitions are explained, serving as a foundation for the assignment. Then, non-parametric testing is elaborated to show the difference between other types of testing and how non-parametric tests are conducted. With this knowledge, two different non-parametric tests are described in detail. This explanation serves to demonstrate the procedure of the methods and how to interpret the respective results. Finally, an experiment featuring a real-life example is conducted, with a non-parametric test. The experiment shows how to apply the method to a practical example. The paradigm and methodology in this written assignment is not expanded on, and no interpretations for future outlooks are made. The purpose of this assignment is only to demonstrate the previously mentioned points. The foundation and inspiration for this assignment is the coursebook “Statistics: Inferential Statistics” (Stefan Stöckl, 2020).

2.1. Framework of statistical hypotheses testing

At first, it is essential to elaborate on crucial definitions of statistical hypotheses testing. As previously mentioned, this part serves as a foundation to understand the key parameters for statistical testing. It is also explained how the results of testing can be interpreted, which is crucial for a successful hypothesis test.

2.1.1 The different paradigms of statistical hypothesis testing

Overall, there are many different paradigms for hypothesis testing. They are based on the modeling of dependent variables with their respective independent variables as parameters. The difference between them is how the statistical model is used. The feud between the different paradigms has been going on since the 19th century. However, there is still no definitive testing paradigm to this day. Every paradigm has its own strengths and weaknesses and is continuously evolving. Also, depending on the use case, a different methodology may be better suited (Jostein Lillestøl, 2014). The decision, which framework to use can be entirely subjective. Currently, the significance-based hypothesis testing is the most commonly used framework (Burnham & Anderson, 2010). That is why I have chosen this methodology for the written assignment. Therefore, the assigment is only concerned with this specific hypothesis testing framework.

2.1.2. The null hypothesis and alternative hypothesis

When it comes to significance-based hypothesis testing, at first, one has to know what hypothesis they want to test and the corresponding null hypothesis. The null hypothesis, defined as H_0, most of the time represents the status quo, which one wants to reject in their testing, through an alternative hypothesis, defined as H_1. If the testing procedure fails to reject the null hypothesis, then there was insufficient evidence to conclude the contrary. In this case, the null hypothesis stays the accepted fact or the default (ThoughtCo, 2020a). If the test can reject the null hypothesis, then there is an argument for an alternative hypothesis, which means that a new hypothesis is preferred instead of the null hypothesis. The alternative hypothesis can become the new default (Cortinhas & Black, 2014). However, there is also the risk of errors when it comes to hypothesis testing. There are two conceptual types of error, which need to be considered during testing. The first kind of error is the type I error, which is the rejection of a true null hypothesis or also known as a false positive. The second kind of error is the type II error, which is the non-rejection of a false null hypothesis or also known as a false negative. A large part of statistical testing revolves around minimizing the probability of errors. Minimizing the error rates is also a way to improve the overall quality of a hypothesis test ("Type I Error and Type II Error - Experimental Errors in Research," 2020b).

2.1.3. Significance level α and the p-value

When it comes to statistical hypotheses testing, an argument for a new default hypothesis can be made if the evidence against the null hypothesis exceeds a certain statistical threshold. This threshold is defined by various factors and must be determined before the actual testing since this can otherwise influence the procedure of hypothesis testing (Salkind & Rasmussen, 2007). This process is also called a "Test of Significance" (Mindrila & Balentyne, 2013). The significance level of the test defines the threshold, also expressed as α. It is the probability of the hypothesis test to reject the null hypothesis when it is true, which consequently means that alpha is also the probability of a type I error. Therefore, if the significance level has reached a value where it is deemed highly unlikely to have occurred under the null hypothesis, the test is statistically significant. This is defined by the expression $p \leq \alpha$. On the other hand, the test is not statistically significant if the following term is the case $p \geq \alpha$ (Cowles & Davis, 1982). In both expressions, the significance level α is compared with the p-value. The p-value is another critical parameter of statistical hypothesis testing. The p-value is a probability, which indicates how unlikely it is that a statistic is observed under the null hypothesis by chance. The smaller the p-value gets, the less likely it is to observe the data sample under the null hypothesis and vice versa (ThoughtCo, 2020b). With this in mind, the term $p \leq \alpha$ expresses that if the p-value lies under the predetermined significance level, an argument for an alternative hypothesis can be made. However, since the p-value is only a probability and no definitive value, the chance still exists that this result happened by chance. Since the general meaning of the p-value is hard to grasp, it is commonly misused. As I mentioned before, if the chance is highly improbable to have

occurred under the null hypothesis, the test is statistically significant. An α of 5% is the generally accepted maximum level to determine statistical significance. This value of α is also highly debated and can vary depending on the area in which the hypothesis test is conducted (Cowles & Davis, 1982).

2.1.4. The false negative rate β and the power value

The value β, just like the value α, expresses the probability for an error. In the case of α, it is the probability to reject the null hypothesis when it is true. In the case of β it is the probability of failing to reject the null hypothesis, even though the null hypothesis is false. This means that β is the probability that a type II error occurs. Unlike α, however, β does not have a value, which it is compared against. β is the false-negative rate used to calculate the power of a statistical hypothesis test, which is another key parameter. The probability that no type II error occurs is defined by the power. Therefore, with the expression 1- β, one can calculate the power of a test. As I mentioned earlier, minimizing error rates is an important part of statistical testing. In an ideal case, α and β would be set to 0, eliminating all kinds of errors. In practice, however, this is not possible for multiple reasons. Reducing α and β usually correlates with increased sample size. An increased sample size, however, makes the study more expensive and complicated. That is why values for α and β are chosen, which are sufficient for the test. A generally accepted value for β is at around 20%, which results in a power of 80%. Like α, this is only a guide value, which can again vary depending on the area in which the hypothesis test is conducted (Banerjee, Chitnis, Jadhav, Bhawalkar, & Chaudhury, 2009). Another reason is the inverse correlation between α and β. Lowering the value of one variable raises the value of the other. That is why maintaining a healthy average is crucial, so that an error type does not inflate too much, reducing the quality of a test (Mudge, Baker, Edge, & Houlahan, 2012).

2.1.5. Interpreting the results of testing

At last comes the most essential part of statistical hypothesis testing, the interpretation of the results. The interpretation of the results defines what the actual conclusion of the test is. Poorly conducted studies can be prone to errors, which lead to wrong conclusions. In the worst case, studies can even be misused purposely. One must consider different fundamental values when interpreting the results. The significance level α, the p-value, and the power of a test are vital leads to identify the quality of a test. The p-value being under or over the significance level is the most crucial part of this since this indicates if the null hypothesis is rejected or failed to reject. The maximum significance level for statistical significance is 5%. It is also essential to identify in what field this study was conducted, as a 5% threshold may not always be sufficient. Furthermore, it is not sufficient to only verify if the p-value is under or over the significance threshold. It is also important to examine the exact p-value. With a significance level of 0.05 a p-value of 0.003 has stronger evidence to reject a null hypothesis than a p-value of 0.049 (Sarmukaddam, 2012). Another critical point is the correct usage of the p-value. The p-value, through its vague explanation, is often misused. It is only the probability that

indicates how unlikely it is that a statistic is observed under the null hypothesis by chance. It is only a probability, not the exact statement that a null hypothesis is rejected or failed to reject (Wasserstein & Lazar, 2016). The sample size also has a major impact on the p-value, where arguably the most manipulation can be made. The reason is that an increase in sample size simultaneously decreases the p-value. This is an enormous problem, especially in combination with big data, since this means that everything can be made statistically significant with enough sample size (Demidenko, 2016). The power of the study also increases with an increase in sample size. The higher the power of a study, the more likely it is to find a difference if one exists. That is why one must examine if a study was not too underpowered to reject the null hypothesis. The power level depends on different factors. However, the highest power is achieved through an increase in sample size. Therefore, one must analyze the sample size carefully, as a lack of justification for the sample size is a common problem (Suresh & Chandrashekara, 2012). The sample size must be large enough for sufficient statistical power and must be low enough to not decrease the p-value too much. However, not only the parameters need to be analyzed. The study must be perceived from a logical standpoint as well. If the null hypothesis is indeed rejected, what is the result of this? This must be analyzed separately. The magnitude is also a hotspot for misinformation, as the phrase “X has an effect on Y” can mean anything, which is commonly used to spread misinformation. It is crucial to identify what kind of effect the test actually has. Is it a major effect/change that could lead to new revolutionary information, or is it something minor that can be ignored? To conclude, there are many components, which need to be considered, be it key values of the study or logical effects outside of it. Because of all these factors, the results of hypothesis tests are a debate in itself (Sarmukaddam, 2012). In addition to the paradigms of significance-based testing, there are also different methods. In the case of this written assignment, the non-parametric methods are being elaborated.

2.2. Non-parametric tests

The test method is the main procedure of the statistical hypothesis test. The method dictates how the data is used and evaluated and has a major impact on the characteristics of a test. In the following chapter, the procedure of non-parametric testing is elaborated. At first, it is explained what exactly non-parametric tests are, which is crucial to understand the concept behind it. In addition, the differences between non-parametric tests and parametric tests are elaborated. Then, the different use cases for the non-parametric tests are shown. At last, two different methods of non-parametric tests are chosen, and the procedures of them are exemplified.

2.2.1. Introduction to non-parametric tests and their difference to parametric tests

In statistical hypothesis testing, non-parametric test methods are methods of statistical analysis, which do not assume anything about the underlying distribution. This makes the non-parametric test the opposite to parametric tests, as the parametric test assumes, that the population follows a certain distribution, which is most of the time a normal distribution. The non-parametric test is often referred

to as a distribution-free test, but that does not necessarily mean that nothing about the distribution is known. Non-parametric tests can also be used when it is known that the population data does not have a normal distribution or if the population data has a specific distribution with unknown parameters (Pearce & Derrick, 2019). Due to the fewer assumptions of non-parametric tests, they have much broader applicability than parametric tests. Other essential aspects are simplicity and robustness. Non-parametric tests are easier to conduct, making them sometimes preferable to parametric tests. Also, through fewer assumptions, the results are more robust, which means non-parametric tests are less prone to errors in the results. These positives, however, come with a trade-off of less statistical power. This means that larger sample sizes are needed to reach the same statistical power as parametric tests. The methodology must also be chosen before the process or after the collection of the data because of the significant influence on the procedure ("Nonparametric Tests," 2017).

2.2.2. Reasons to use non-parametric tests

When the conditions are not met, non-parametric tests give the right results, whereas parametric tests do not. However, when the normality assumptions match, then parametric tests are more efficient. Therefore, non-parametric tests serve as an alternative to parametric tests (Kaur & Kumar, 2015). The assumptions about the population sample are not met when the data does not behave like a normal distribution. But also, when the data sample behaves like a normal population, skewness, for example, can distort the distribution. This makes the mean not as effective. Generally speaking, non-parametric tests are more effective with distributions, where the median is preferable to the mean. There are different reasons to choose non-parametric tests as well. The first being that the population sample size is too small. This may sound contradicting with the previous statement that non-parametric tests need more sample data to reach the same statistical power. However, parametric methods can only be applied if the sample size is large enough to determine the distribution. The determination of the distribution might not be possible if the sample size is too small. (Institute, Corporate Finance, 2020). Another reason being the type of data. While parametric tests always use quantitative data, non-parametric tests can also be used for ordinal and nominal data. These are the level of measurements, which do not have a clear numerical value. Labels, categories, and rankings, for example, fall under this classification (McHugh, 2003).

2.2.3. Mann-Whitney U-test

The first non-parametric test method elaborated is the Mann-Whitney U-test developed by Mann, Whitney, and Wilcoxon. The test analyses, if two independent data samples represent the same population or two different populations with different median values (MacFarland & Yates, 2016). For the Mann-Whitney U-test, ordinal data is employed. This means data with rank-order. An interval/ratio score can also be transformed into an ordinal score if deemed appropriate. The Mann-Whitney U-test is used to determine if there are statistically significant differences between two independent

samples. If the result of the test is statistically significant, it indicates a difference between the medians, which leads to the conclusion of two different populations. The procedure of the method is conducted as follows. There is a total number of subjects when conducting the test, which is denoted by n. For example, there are 10 subjects. These 10 subjects are then randomly divided into groups, each with 5 subjects. These groups are “Group 1” and “Group 2”. One group acts as a control group, while the other acts as the experiment group. The control group is here to be the default or, in this case, the null hypothesis, while the experiment group is the effect or the alternative hypothesis. Every subject in the test has a identification, which is a unique labeling. Based on what kind of test is performed, the test is assigned a ranking, representing the magnitude of what is tested—for example, the ranking of depression as a test. “Group 1” is prescribed a newly developed drug, while “Group 2” is prescribed a placebo. After an extended period of time, the results are collected and distributed into ranks. The following table shows the scores and complementary ranks of the subjects.

	Group 1			Group 2	
	X_1	R_1		X_2	R_2
Subject 1,1	11	9	**Subject 1,2**	11	9
Subject 2,1	1	3	**Subject 2,2**	11	9
Subject 3,1	0	1.5	**Subject 3,2**	5	6
Subject 4,1	2	4	**Subject 4,2**	8	7
Subject 5,1	0	1.5	**Subject 5,2**	4	5
		$\Sigma R_1 = 19$			$\Sigma R_2 = 36$
	$\bar{R}_1 = \frac{\Sigma R_1}{n_1} = \frac{19}{5} = 3.8$			$\bar{R}_2 = \frac{\Sigma R_2}{n_2} = \frac{36}{5} = 7.2$	

Table 1. Table of the score to rank conversion for the Mann-Whitney U-test (Sheskin, 2003)

The subjects are then ranked in order and the tied ranks are adjusted with the formula $\frac{n_1+n_2+\cdots n_x}{n}$.

Subject identification number	3,1	5,1	2,1	4,1	5,2	3,2	4,2	1,1	1,2	2,2
Depression score	0	0	1	2	4	5	8	11	11	11
Rank prior to tie adjustment	1	2	3	4	5	6	7	8	9	10
Tie-adjusted rank	1.5	1.5	3	4	5	6	7	9	9	9

Table 2. Table of the score order and tie rank adjustment for the Mann-Whitney U-test (Sheskin, 2003)

Since all of the subjects have been assigned a rank, and the sum of the ranks was computed, the test statistic U can be computed. U is calculated as follows, for “Group 1” $U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1+1)}{2} - \sum R_1$ and for “Group 2” $U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2+1)}{2} - \sum R_2$. When the correct values are inserted, the results are $U_1 = 21$ and $U_2 = 4$. A way of confirmation for a successful application is the formula $n_1 \cdot n_2 = U_1 + U_2$

and the non-existence of negative values. Since both conditions are met, a correct application can be assumed. The value of U is then evaluated with the "Table of Critical Values for the Mann-Whitney U Statistics", which is in the Appendix. The tables have variations based on the tail of the test and the significance level. Where the rows and columns of n_1 and n_2 meet is the critical value for the test. For the test to be significant, the obtained U value must be equal to or less than the critical value. The value, which is compared with the critical value, is the smallest U, which would be $U_2 = 4$ in this case. Since this is the value, which must be employed to reach the critical value, the alternative hypothesis must be stated as $H_1: \theta_1 < \theta_2$. For this alternative hypothesis to be supported, the average of ranks in "Group 2" must be higher than the average of the ranks in "Group 1". The alternative hypothesis is supported at a 0.05 significance level since U = 4 is equal with the one-tailed value $U_{.05} = 4$. The results in this case state that the drug lowered the degree of depression in the experimental group, with a statistical significance of 5% (Sheskin, 2003).

2.2.4. The McNemar Test

The second non-parametric test method elaborated is the McNemar Test. For the McNemar Test, categorical data, also known as nominal data, is used. The testing involves a contingency table with 2 x 2 size and n samples, possessing a dichotomous trait. The dichotomous trait describes a whole set, divided into parts. All the samples in this table are jointly exhaustive, meaning every sample belongs to one of the parts, and mutually exclusive, meaning no sample exists simultaneously in the different parts. The McNemar test is employed to examine the marginal homogeneity of the table. This is the case if row and column marginal frequencies are equal. This serves to check if two experimental samples come from two different populations or if there is a difference between pretest and posttest scores of samples on the dependent variable. The McNemar test is based on the following assumptions, which must be met for the test to be legit. At first, the n subjects have been randomly selected and are independent of the other observations. Then, the scores of subjects must follow the rules of the dichotomous trait. And at last, the test should not be conducted with very small sample sizes. The test is conducted as follows. The cells contain the scores of subjects, which are represented by two themes and four categories, which shapes the 2 x 2 table. Therefore, every margin is a total of two different cells, each representing different combinations. The following example describes the testing of a drug and a placebo.

		Favorable response to drug		
		Yes	No	Row sums
Favorable response to placebo	Yes	10	13	23
	No	41	36	77
	Column sums	51	49	100

Table 3. 2x2 table of the medical survey for the McNemar Test (Sheskin, 2003)

After the scores are collected, it is important to identify the key parameters of the test. In this case, they would be Cell b = 13 and Cell c = 41. Cell b is the favorable response to the placebo/unfavorable response to the drug, and Cell c is the unfavorable response to the placebo/favorable response to the drug. One wants to test if the drug is more effective than the placebo, which would mean that the score supporting the drug needs to be larger than the score supporting the placebo. Regarding the underlying populations, p_b and p_c represent the following proportions $P_b = \frac{b}{(b+c)}$ and $P_c = \frac{c}{(b+c)}$. When the values from the experiment are inserted the results are $p_b = 0.24$ and $p_c = 0.76$. If there is no difference between the null hypothesis and the alternative hypothesis, the following will be true $p_b = p_c = 0.5$. Since $p_c > p_b$ there is an argument for an effect of the drug. The alternative hypothesis can be stated as H_1: $p_c > p_b$. The test statistic is then calculated with the equation $\chi^2 = \frac{(b-c)^2}{b+c}$. When the values of the experiment are inserted, the result is $\chi^2 = 14{,}52$. The computed value must always be positive, since a negative number indicates and error. The obtained value is then evaluated with the "Table of Chi-Square Distribution", which is in the Appendix. For this experiment, the degress of freedom (df) = 1, because of the 2 x 2 table, and the significance level is 1%. Since the obtained value of χ^2 is greater than the $\chi^2_{.01} = 6{,}63$, the experiment has statistical significance. In the context of this experiment this would conclude that a greater proportion of subjects respond significantly more favourable to the drug than to the placebo (Sheskin, 2003).

2.3. Conducting the experiment

At last, a statistical hypothesis test is conducted on a real-life example using one of the previously elaborated non-parametric test methods. The first step here is the elaboration of the approach. Here, the topic of the experiment is introduced, and preparations, which need to be done beforehand, are explained. Then, the non-parametric test method is applied and elaborated, why this method is suited for this experiment. At last, the results are discussed and interpreted.

2.3.1. Elaboration of the approach

The most crucial step at the beginning of the approach is the stating of the hypothesis. A hypothesis should be as clearly defined as possible since this can otherwise impair the overall experiment ("Statistical Hypothesis: Definition," 2020a). The hypothesis for this experiment is "The 20-30-year-old Austrian population value free time more than money, compared to the 30-40-year-old Austrian population". In this experiment, randomly selected individuals from their respective age group are questioned about their perspective. The exact question to the individuals is "Do you value free time more than money?" which is represented by the Likert scale. The Likert scale is a rating scale of agreement and disagreement (McKelvie, 2015). The Likert scale in this experiment consists of five categories, with their own score.

Disagree = 1, Slightly Disagree = 2, Neutral = 3, Slightly Agree = 4, Agree = 5.

2.3.2. Applying a non-parametric test on the experiment

Once the foundation for the experiment has been decided, the approach must be thought through. The non-parametric test chosen for this experiment is the Mann-Whitney U-test since this is a perfect fit to test two groups against each other. Then the number of samples must be chosen for the experiment. For a non-parametric test, there is no direct way to compute the required sample size, and the "correct" size of samples is highly debatable. There are multiple possibilities for choosing the sample size. The approach taken here is a rough estimate of the needed sample size. The reasoning is the not available parameters for a more accurate computation and the type of data used for the test. Non-parametric tests also need more sample size to compensate for the lack of power, contrary to their parametric counterpart. For a sufficient sample size, the requirements of samples per group when conducting the 2-sample t-test is about 15. A non-parametric counterpart needs approximately 15% more samples to counteract the lack of power. A good sample size for this test is about 20 subjects per group (Frost, 2017). After the number of samples have been decided, and the survey has been conducted, the key parameters are inserted into the table. This looks like the following.

20-30 year old austrian population			**30-40 year old austrian population**		
Subjects	Scores	R_1	**Subjects**	Scores	R_2
Subject 1,1	5	35	**Subject 1,2**	1	4,5
Subject 2,1	4	25,5	**Subject 2,2**	2	12
Subject 3,1	5	35	**Subject 3,2**	1	4,5
Subject 4,1	5	35	**Subject 4,2**	3	18,5
Subject 5,1	3	18,5	**Subject 5,2**	4	25,5
Subject 6,1	4	25,5	**Subject 6,2**	2	12
Subject 7,1	5	35	**Subject 7,2**	5	35
Subject 8,1	2	12	**Subject 8,2**	1	4,5
Subject 9,1	1	4,5	**Subject 9,2**	2	12
Subject 10,1	3	18,5	**Subject 10,2**	4	25,5
Subject 11,1	5	35	**Subject 11,2**	5	35
Subject 12,1	5	35	**Subject 12,2**	2	12
Subject 13,1	5	35	**Subject 13,2**	4	25,5
Subject 14,1	4	25,5	**Subject 14,2**	3	18,5
Subject 15,1	1	4,5	**Subject 15,2**	1	4,5
Subject 16,1	5	35	**Subject 16,2**	1	4,5
Subject 17,1	4	25,5	**Subject 17,2**	2	12
Subject 18,1	2	12	**Subject 18,2**	3	18,5
Subject 19,1	3	18,5	**Subject 19,2**	1	4,5
Subject 20,1	4	25,5	**Subject 20,2**	5	35
$\sum \mathbf{R_1}$ = 496 and $\overline{\mathbf{R}}_{\mathbf{1}} = \frac{\mathbf{496}}{\mathbf{20}}$ = 24,8			$\sum R_2$ =324 and $\overline{R}_2 = \frac{324}{20}$ = 16,2		

Table 4. Score to rank conversion of the experiment

The subjects are then ranked in order of their scores and the tied ranks are adjusted.

Subject number	9,1	15,1	1,2	3,2	8,2	15,2	16,2	19,2	8,1	18,1	2,2
Score	1	1	1	1	1	1	1	1	2	2	2
Rank prior	1	2	3	4	5	6	7	8	9	10	11
Adjusted rank	4,5	4,5	4,5	4,5	4,5	4,5	4,5	4,5	12	12	12

Subject number	6,2	9,2	12,2	17,2	5,1	10,1	19,1	4,2	14,2	18,2	2,1
Score	2	2	2	2	3	3	3	3	3	3	4
Rank prior	12	13	14	15	16	17	18	19	20	21	22
Adjusted rank	12	12	12	12	18,5	18,5	18,5	18,5	18,5	18,5	25,5
Subject number	6,1	14,1	17,1	20,1	5,2	10,2	13,2	1,1	3,1	4,1	7,1
Score	4	4	4	4	4	4	4	5	5	5	5
Rank prior	23	24	25	26	27	28	29	30	31	32	33
Adjusted rank	25,5	25,5	25,5	25,5	25,5	25,5	25,5	35	35	35	35
Subject number	11,1	12,1	13,1	16,1	7,2	11,2	20,2				
Score	5	5	5	5	5	5	5				
Rank prior	34	35	36	37	38	39	40				
Adjusted rank	35	35	35	35	35	35	35				

Table 5. Score order and tie rank adjustment of the experiment

After all the ranks are adjusted and the total is calculated, the U values from the different groups are calculated, with the formulas mentioned previously. After inserting the values, the results are U_1 = 114 and U_2 = 286. Since the U values are not negative and the formula $n_1 \cdot n_2 = U_1 + U_2$ applies, a correct application can be assumed. For this experiment, the ranks reflect the importance of free time over money. That is why the hypothesis needs to be stated as $H_1: \theta_1 > \theta_2$. The alternative hypothesis is tested in a one-tailed test, at a 0.01 significance level. The U from the experiment is 114, and since the $U_{.01}$ at 20 samples for each group equals 114, the alternative hypothesis is supported. The results are that the 20-30-year-old Austrian population values free time more than money, compared to the 30-40-year-old Austrian population at a 1% significance level. The effect size, however, can not be properly determined since there are no specific percentages. The effect size could only be measured indirectly on different areas, which this mindset impacts.

2.4. Conclusion

The overall area of hypothesis testing is broad and convoluted. The different paradigms, methodologies, and uncountable models make the decision, even what to use, a complicated one. As I stated in the beginning, the reason for choosing the significance-based testing is because it is the most common used method, but it is also the most criticized. The p-value, the choice of sample sizes, the easy manipulation, the significance level, and the paradigm as a whole have been subject of criticism for a long time now. Stating the effect sizes, the newly interpreted p-value and additional parameters is a new step to counteract the common misinterpretation and misuses. The foundation of the paradigm and especially the p-value is viewed by many as flawed. Nonetheless, it is still the most used paradigm and is continually being refined. However, the other schools of hypothesis testing are also refined to this day, and researchers have been switching to the Bayes paradigm as well. All in all, it is difficult to say how the hypothesis testing paradigms will evolve and what kind of paradigm will prevail.

III. List of Appendices

IV. Appendices

Appendix 1. Table of Critical Values for Mann-Whitney U-Statistic (One-tailed .05 Values)

(Sheskin, 2003)

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																			0	0
2					0	0	0	1	1	1	1	2	2	2	3	3	3	4	4	4
3			0	0	1	2	2	3	3	4	5	5	6	7	7	8	9	9	10	11
4			0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
5		0	1	2	4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
6		0	2	3	5	7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
7		0	2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37	39
8		1	3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44	47
9		1	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54
10		1	4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	55	58	62
11		1	5	8	12	16	19	23	27	31	34	38	42	46	50	54	57	61	65	69
12		2	5	9	13	17	21	26	30	34	38	42	47	51	55	60	64	68	72	77
13		2	6	10	15	19	24	28	33	37	42	47	51	56	61	65	70	75	80	84
14		2	7	11	16	21	26	31	36	41	46	51	56	61	66	71	77	82	87	92
15		3	7	12	18	23	28	33	39	44	50	55	61	66	72	77	83	88	94	100
16		3	8	14	19	25	30	36	42	48	54	60	65	71	77	83	89	95	101	107
17		3	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96	102	109	115
18		4	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109	116	123
19	0	4	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	130
20	0	4	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138

Appendix 2. Table of Critical Values for Mann-Whitney U-Statistic (One-tailed .01 Values)

(Sheskin, 2003)

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2													0	0	0	0	0	0	1	1
3							0	0	1	1	1	2	2	2	3	3	4	4	4	5
4					0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10
5				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
6				1	2	3	4	6	7	8	9	11	12	13	15	16	18	19	20	22
7			0	1	3	4	6	7	9	11	12	14	16	17	19	21	23	24	26	28
8			0	2	4	6	7	9	11	13	15	17	20	22	24	26	28	30	32	34
9			1	3	5	7	9	11	14	16	18	21	23	26	28	31	33	36	38	40
10			1	3	6	8	11	13	16	19	22	24	27	30	33	36	38	41	44	47
11			1	4	7	9	12	15	18	22	25	28	31	34	37	41	44	47	50	53
12			2	5	8	11	14	17	21	24	28	31	35	38	42	46	49	53	56	60
13		0	2	5	9	12	16	20	23	27	31	35	39	43	47	51	55	59	63	67
14		0	2	6	10	13	17	22	26	30	34	38	43	47	51	56	60	65	69	73
15		0	3	7	11	15	19	24	28	33	37	42	47	51	56	61	66	70	75	80
16		0	3	7	12	16	21	26	31	36	41	46	51	56	61	66	71	76	82	87
17		0	4	8	13	18	23	28	33	38	44	49	55	60	66	71	77	82	88	93
18		0	4	9	14	19	24	30	36	41	47	53	59	65	70	76	82	88	94	100
19		1	4	9	15	20	26	32	38	44	50	56	63	69	75	82	88	94	101	107
20		1	5	10	16	22	28	34	40	47	53	60	67	73	80	87	93	100	107	114

Appendix 3. Table of the Chi-Square Disitribution (Sheskin, 2003)

p	.005	.010	.025	.050	.100	.900	.950	.975	.990	.995	.999
df											
1	.0393	.0157	.0982	.0393	.0158	2.71	3.84	5.02	6.63	7.88	10.83
2	.0100	.0201	.0506	.103	.211	4.61	5.99	7.38	9.21	10.60	13.82
3	.072	.115	.216	.352	.584	6.25	7.81	9.35	11.34	12.84	16.27
4	.0207	.297	.484	.711	1.064	7.78	9.49	11.14	13.28	14.86	18.47
5	.412	.554	.831	1.145	1.61	9.24	11.07	12.83	15.09	16.75	20.52
6	.676	.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55	22.46
7	.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96	26.13
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80	37.70
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27	39.25
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72	40.79
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16	42.31
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58	43.82
20	7.43	8.26	8.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00	43.32
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40	46.80
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80	48.27
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18	49.73
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99	56.89
29	13.21	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67	59.70
40	20.71	22.16	24.43	26.51	29.05	51.80	55.76	59.34	63.69	66.77	73.40
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49	86.66
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95	99.61
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43	104.22	112.32
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	116.32	124.84
90	59.20	61.75	65.65	69.13	73.29	107.56	113.15	118.14	124.12	128.30	137.21
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81	140.17	149.45

V. Bibliography

Banerjee, A., Chitnis, U. B., Jadhav, S. L., Bhawalkar, J. S., & Chaudhury, S. (2009). Hypothesis testing, type I and type II errors. *Industrial Psychiatry Journal*, *18*(2), 127. https://doi.org/10.4103/0972-6748.62274

Burnham, K. P., & Anderson, D. R. (2010). *Model selection and multimodel inference: A practical information-theoretic approach* (2. ed.). New York, NY: Springer.

Cortinhas, C., & Black, K. (2014). *Statistics for Business and Economics*. Hoboken: Wiley Textbooks. Retrieved from http://gbv.eblib.com/patron/FullRecord.aspx?p=882724

Cowles, M., & Davis, C. (1982). On the origins of the.05 level of statistical significance. *American Psychologist*, *37*(5), 553–558. https://doi.org/10.1037/0003-066X.37.5.553

Demidenko, E. (2016). The p-Value You Can't Buy. *The American Statistician*, *70*(1), 33–38. https://doi.org/10.1080/00031305.2015.1069760

Frost, J. (2017, April 11). Nonparametric Tests vs. Parametric Tests. *Jim Frost*. Retrieved from https://statisticsbyjim.com/hypothesis-testing/nonparametric-parametric-tests/

Institute, Corporate Finance (2020, May 24). Nonparametric Tests. *Corporate Finance Institute*. Retrieved from https://corporatefinanceinstitute.com/resources/knowledge/other/nonparametric-tests/

Jostein Lillestøl (2014). Statistical Interference: Paradigms and controversies in historic perspective. Retrieved from https://www.nhh.no/globalassets/departments/business-and-management-science/research/lillestol/statistical_inference.pdf

Kaur & Kumar (2015). Comparative Analysis of Parametric and Non-Parametric Tests. Retrieved from http://dsc.du.ac.in/wp-content/uploads/2020/04/lecture-13-parametric-and-non-parametric-tests.pdf

MacFarland, T. W., & Yates, J. M. (Eds.) (2016). *Introduction to Nonparametric Statistics for the Biological Sciences Using R*. Cham: Springer. https://doi.org/10.1007/978-3-319-30634-6

McHugh, M. L. (2003). Descriptive statistics, Part I: Level of measurement. *Journal for Specialists in Pediatric Nursing : JSPN*, *8*(1), 35–37. https://doi.org/10.1111/j.1744-6155.2003.tb00182.x

McKelvie (2015, May 30). What is a Likert Scale? Retrieved from http://core.ecu.edu/psyc/wuenschk/StatHelp/Likert.htm

Mindrila, & Balentyne (2013). Tests of Significance. Retrieved from https://www.westga.edu/academics/research/vrc/assets/docs/tests_of_significance_notes.pdf

Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an optimal α that minimizes errors in null hypothesis significance tests. *PloS One*, *7*(2), e32734. https://doi.org/10.1371/journal.pone.0032734

Nonparametric Tests (2017, May 4). Retrieved from https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/bs704_nonparametric_print.html

Pearce, J., & Derrick, B. (2019). Preliminary Testing: The Devil of Statistics? *Reinvention: An International Journal of Undergraduate Research*, *12*(2). https://doi.org/10.31273/reinvention.v12i2.339

Salkind, N. J., & Rasmussen, K. (2007). *Encyclopedia of measurement and statistics*. Thousand Oaks, Calif.: SAGE Publications. Retrieved from http://site.ebrary.com/lib/uniregensburg/Doc?id=10367430

Sarmukaddam, S. B. (2012). Interpreting "statistical hypothesis testing" results in clinical research. *Journal of Ayurveda and Integrative Medicine*, *3*(2), 65–69. https://doi.org/10.4103/0975-9476.96518

Sheskin, D. J. (2003). *Handbook of Parametric and Nonparametric Statistical Procedures*: Chapman and Hall/CRC. https://doi.org/10.1201/9781420036268

Statistical Hypothesis: Definition (2020a, November 21). Retrieved from https://stattrek.com/statistics/dictionary.aspx?definition=statistical-hypothesis

Stefan Stöckl (2020, November 17). IUBH Reader - Webreader.io. Retrieved from https://iubh.webreader.io/#!/reader/025e3f7b-d4f5-4577-8ac4-91e6375295f6/page/e4bce09d-8d28-4bd4-a23e-1efd0a0ff13a

Suresh, K., & Chandrashekara, S. (2012). Sample size estimation and power analysis for clinical research studies. *Journal of Human Reproductive Sciences*, *5*(1), 7–13. https://doi.org/10.4103/0974-1208.97779

ThoughtCo (2020a, November 23). What Is the Null Hypothesis? Definition and Examples. Retrieved from https://www.thoughtco.com/definition-of-null-hypothesis-and-examples-605436

ThoughtCo (2020b, December 1). What Is a P-Value. Retrieved from https://www.thoughtco.com/what-is-a-p-value-3126392

Type I Error and Type II Error - Experimental Errors in Research (2020b, November 26). Retrieved from https://explorable.com/type-i-error

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108