Illinois State University
ISU ReD: Research and eData

Theses and Dissertations

2021

Proposing Alternative Methods for Testing Heterogeneity of Studies in a Meta-analysis

Teagen Smith
Illinois State University, teagens1017@gmail.com

Follow this and additional works at: https://ir.library.illinoisstate.edu/etd

Recommended Citation
Smith, Teagen, "Proposing Alternative Methods for Testing Heterogeneity of Studies in a Meta-analysis" (2021). *Theses and Dissertations*. 1471.
https://ir.library.illinoisstate.edu/etd/1471

This Thesis-Open Access is brought to you for free and open access by ISU ReD: Research and eData. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of ISU ReD: Research and eData. For more information, please contact ISUReD@ilstu.edu.

PROPOSING ALTERNATIVE METHODS FOR TESTING HETEROGENEITY OF STUDIES IN A META-ANALYSIS

TEAGEN SMITH

59 Pages

A meta-analysis is a tool commonly used to try and gain an understanding of a given topic by using multiple studies conducted on the topic. A key element of properly interpreting the results of a meta-analysis is the test to check for heterogeneity within the studies included. This is currently done using Cochrane's Q-statistic to test a null hypothesis that the studies included share a common effect size. However, this method has been scrutinized for some of its downsides such as its low power in cases with small sample sizes. This can often create issues because meta-analysis is commonly used in scenarios in which there are few studies. Therefore, in this paper we decide to propose some alternative methods for testing heterogeneity among the studies of a meta-analysis. These methods include using U-statistics because of a few helpful characteristics that may make their interpretation with regard to meta-analysis simpler. It is important to note that meta-analysis is a widely used tool in research and is not limited to use in statistical fields. Therefore, the interpretability of the statistic used to test the heterogeneity is crucial in some cases and we aim to make this easier by recommending an alternative to Cochrane's Q. This proposed method, while simple, will hopefully lay the groundwork for easily interpretable and more accurate tests for heterogeneity using U-statistics.

KEYWORDS: meta-analysis, heterogeneity, Cockrane's Q, U-statistics

PROPOSING ALTERNATIVE METHODS FOR TESTING HETEROGENEITY OF STUDIES IN A META-ANALYSIS

TEAGEN SMITH

A Thesis Submitted in Partial
Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

Department of Mathematics

ILLINOIS STATE UNIVERSITY

2021

© 2021 Teagen Smith

PROPOSING ALTERNATIVE METHODS FOR TESTING HETEROGENEITY OF STUDIES IN A META-ANALYSIS

TEAGEN SMITH

COMMITTEE MEMBERS:

Pei Geng, Chair

Olcay Akman

John Sedbrook

CONTENTS

TABLES

FIGURES

CHAPTER I: INTRODUCTION

I.1 Meta-analysis Background

Often times, multiple studies are conducted to understand important or difficult questions. One way to use the information from all of these studies is to conduct a meta-analysis. A meta-analysis is a statistical method in which the results of multiple studies are combined in an attempt to draw an overall conclusion based on the findings of these studies. Some believe that the use of meta-analyses began in 1904 when used by Pearson in a study that pooled the results of clinical trials of typhoid vaccinations (Jones, 1995). However, after that point in time this type of analysis was increasingly used in the field of psychology with use in science and medical fields becoming more common later on, and the term meta-analysis was officially introduced in 1976. (Jones, 1995). The prefix meta is used to mean comprehensive and analysis is used to understand a given topic. Therefore, this word is used to mean conducting a comprehensive analysis or gaining a comprehensive understanding of a given topic. The process of conducting a meta-analysis can be divided into three basic parts.

The first part of a meta-analysis is creating the research question (Page et al., 2021, Borenstein et al., 2009). The type of research question will ultimately impact the decisions made in the following steps of the meta-analysis. For example, a meta-analysis focusing on which medication is better for a given disease could be conducted differently than one focused on understanding mating times of a given species. Thus, it is important to put thought into the initial research question and what exactly one hopes to gain from answering this question. The next part of conducting a meta-analysis involves searching through the literature and selecting studies to include in the meta-analysis (Jones et al., 2008). Although, this may sound relatively simple, consider the large amount of studies produced for some topics. This next step includes systematically searching through the literature to find articles that seem relevant to the chosen topic. The criteria for selecting studies for a meta-analysis should be well defined so that the studies used are as similar as possible and the criteria should be defined before searching

through the literature (Jones et al., 2008). Then, articles that do not meet specific criteria are removed from the analysis. The data is then retrieved from the studies that do meet the criteria. It is important to note that it is best that multiple authors decide on which studies are included in the study to eliminate bias (Jones et al., 2008). Finally, the last part of a meta-analysis includes synthesizing the data pulled from each study and attempting to draw conclusions. The following sections will discuss this part of a meta-analysis in more detail. Generally, this includes calculating effect sizes for each study, finding a summary effect using a fixed or random effects model, checking the heterogeneity of the studies, and drawing conclusions on the results of the meta-analysis. Understanding how to properly implement each of these parts of the meta-analysis is crucial in properly interpreting the results.

As stated previously, checking the heterogeneity of the studies included in the meta-analysis is an important step, therefore we give a formal definition with regard to meta-analysis as well as an example. Heterogeneity is defined as the differences or variability between studies included in a meta-analysis other than differences that could be explained by random error (Kolasa & Rollo, 1991). Conversely, we would define homogeneity as studies where the only differences that are observed can be contributed to random error. It is important to identify heterogeneity between studies in a meta-analysis because if it is present, then general conclusions cannot be drawn from the combined results of the studies. For example, consider a meta-analysis that consists of studies conducted on the effectiveness of a gasoline additive. If about half of the studies showed that this additive increased gas milage and half showed it decreased milage, then we would more than likely see that heterogeneity exists between studies. Intuitively, we would not try to draw a general conclusion based on these studies because they show such different results. Alternatively, consider the scenario in which all studies showed a similar improvement in gas milage. In this case, we would more than likely not see heterogeneity between studies and may even be able to conclude something about the effectiveness of

this gasoline additive. Therefore, it is essential that one accurately tests for heterogeneity between studies included in a meta-analysis.

The current method for checking heterogeneity has been criticized. Therefore, the goal of this paper is to recommend an alternative method to test for heterogeneity in meta-analyses. To do this, one must completely understand the steps of a meta-analysis, thus we explain the analytical steps in more detail. After this, we give some background on U-statistics which is the statistic we propose using instead of Cochrane's Q, which is currently used. Then examples of how this new test would be conducted are given and compared to the current test. Finally, we discuss how it compares and what this could mean for future research in this area.

It is important to note that the methods proposed in this paper could be very beneficial in the application of meta-analysis in a variety of fields, specifically in the field of biomathematics. One reason being that in the field of biomathematics, meta-analysis is a tool that is used fairly often. Consider the COVID-19 pandemic and the many studies have been released pertaining to the pandemic. Meta-analysis has been used to synthesize these studies and to help understand the virus. In this case, it is essential that researchers properly test for heterogeneity among the studies included so that no incorrect conclusions are drawn. These proposed methods for testing heterogeneity have the potential to be important in the field of biomathematics outside of their use in meta-analysis. It is a well-known fact that the existence of heterogeneity plays a crucial role in the modeling of biological systems. In the case of this paper, we are only concerned with statistical heterogeneity, however there are other types such as spatial, clinical, or methodological heterogeneity that can impact the way that a researcher models a biological system (Fletcher, 2007). Because U-statistics are a class of statistics that are commonly used for testing, the proposed statistics have the potential to test for the existence other types of heterogeneity. Therefore, the methods proposed in this paper may be especially helpful in the field of biomathematical modeling. Throughout this paper, we explain how to properly implement the

proposed U-statistics while also keeping in mind how they are used and interpreted in fields like biomathematics.

I.2 Calculating Effect Sizes

For each of the studies included in the meta-analysis, an effect size must be found. The effect size is also referred to as the treatment effect (as well as other names depending on the field in which the meta-analysis is based on or its' goal). Thus, some consideration must go into the selection of the type of effect size calculation that will be used. One should note that the effect size needs to be comparable between studies (so it should not depend on factors such as sample size), it should use information from the raw data of the studies, and it should be relatively easy to work with in a way that variance and confidence intervals are able to be found (Borenstein et al., 2009). A few types of effect sizes that are used often are raw mean difference, standardized mean difference, response ratios, risk ratios, odds ratios, risk difference, and correlations (Borenstein et al., 2009). Again, the selection of effect size measure is dependent on the type of data from the studies as well as the field, and the previous list is not exhaustive. For the previous list however, we do know how to calculate the variance, standard error, and thus the confidence intervals. In the following sections, we focus on the raw mean difference and standardized mean difference because they are some of the more common methods to calculate effect size. Also, in the worked examples later in this paper, we only use the standardized mean difference to find the effect sizes. The true effect size is denoted as the parameter θ; however, we denote our estimate of the effect size of the study with Y.

Raw Mean Difference

The raw mean difference is used best in a scenario where all of the studies are using the same scale (Borenstein et al., 2009). Consider a measure for a study in which the actual value is important such a study involving calories. We would want to keep the scale the same in this case. Intuitively, this method makes the most sense as it is just the difference between the two groups being considered in a

study. This is an estimate of the true difference in the population means for a study. Thus, the raw difference is calculated as follows:

$$D = \overline{X_1} - \overline{X_2} \tag{1.1}$$

Note that we are using two sample means here. To calculate this effect size, we are using two groups to interpret the effect in each study. Then for the calculation of variance and standard deviation we either assume the populations standard deviations are the same or they are not. If we assume that the standard deviations of the populations are the same then we calculate the variance for the raw mean differences to be

$$V_D = \frac{n_1 + n_2}{n_1 n_2} S^2_{pooled} \tag{1.2}$$

such that

$$S_{pooled} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \tag{1.3}$$

Otherwise, we assume the standard deviations are different and we calculate the variance for the raw mean differences as follows

$$V_D = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \tag{1.4}$$

For both of these cases we take S_1 and S_2 to be the sample standard deviations of the two groups and n_1 and n_2 are the sample population sizes of the two groups. And obviously in both cases we have the standard error to be the square root of the estimated variance

$$SE_D = \sqrt{V_D} \tag{1.5}$$

Standardized Mean Difference

The standardized mean difference is often used in scenarios where the raw mean difference is not appropriate. For example, using the standardized mean difference is better for comparing studies that use different instruments from one another that could cause inconsistencies (Borenstein et al.,

2009). Similar to the raw mean difference, this method is calculating the effect size between two groups. Thus, we estimate the standardized mean difference between two groups as follows

$$d = \frac{\overline{X_1} - \overline{X_2}}{S_{within}} \tag{1.6}$$

where the numerator is the difference between the two groups sample means. Then the denominator is the within groups standard deviation and is calculated as

$$S_{within} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \tag{1.7}$$

which is the same as the S_{pooled} value that was calculated for the raw mean difference above. Again, for this value we use the sample sizes and sample standard deviations from the two groups that we are concerned with. The estimate of the effect size, d, here is often referred to as Cohen's d (Borenstein et al., 2009). Now that we have this value, we would calculate the estimated variance as follows

$$V_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)} \tag{1.8}$$

And as previously, we calculate the estimates standard error as the square root of the estimated variance

$$SE_D = \sqrt{V_D} \tag{1.9}$$

After calculating the effect sizes and the variances for each of the studies, we then use the values in the next few steps of conducting a meta-analysis. For both the fixed effects and random effects models, we use these values in estimating the summary effect for the studies. Later in the process, we also use these values in the computations to test for heterogeneity between studies.

I.3 Fixed Effect vs Random Effect Models

Fixed Effects Model

The main assumption of the fixed effects model is that all of the studies included in the meta-analysis share a common effect size. This is referring to the true effect size of the studies, which is

typically referred to as θ and is unknown (Borenstein et al., 2009). Therefore, this model assumes that the effect size is calculated as

$$Y_i = \theta + \varepsilon_i \tag{1.10}$$

which implies that the only way that studies would vary from each other would be through random error. Although this model assumes that the studies all share the same true effect size, they can differ in sample size or error and for that reason are given weights. The weight of each study in a fixed effects meta-analysis is calculated as

$$W_i = \frac{1}{V_{Y_i}} \tag{1.11}$$

where just as before V_{Y_i} is the variance of the studies estimated effect size. We are then able to use these weights to calculate the estimated summary effect in such a way that studies with small sample sizes are not over-represented. The weighted mean or the summary effect is then

$$M = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i} \tag{1.12}$$

We can also estimate a variance for the summary effect in a way similar to that for a single study's effect size. Thus, we intuitively calculate this as

$$V_M = \frac{1}{\sum_{i=1}^{k} W_i} \tag{1.13}$$

and obviously an estimated standard error for the summary effect is calculated as the square root of the estimated variance.

$$SE_M = \sqrt{V_M} \tag{1.14}$$

We know that the goal of using the model is to estimate the summary effect for the group of studies that have been selected for the meta-analysis. Therefore, a few more calculations are used to help us understand how good the estimate is. So, next we consider the 95% upper and lower confidence limits which are calculated as

$$LL_M = M - 1.96 \times SE_M \tag{1.15}$$

$$UL_M = M + 1.96 \times SE_M \tag{1.16}$$

Using these, we gain understanding on the summary effect estimate. We obviously hope to not have a very large confidence interval. Another value we calculate is the Z-value for our summary effect. We calculate the Z-value as

$$Z = \frac{M}{SE_M} \tag{1.17}$$

And then we calculate a p-value to test the hypothesis that that the true effect size is zero (Borenstein et al., 2009). A significant p-value would lead us to reject the null hypothesis and conclude that the true effect size is not zero. In other words, this would mean that the treatment of interest does truly have an effect. We calculate the one-tailed p-value as

$$p = 1 - \Phi(\pm|Z|) \tag{1.18}$$

And in the case that a two-tailed p-value is used, it is calculated as

$$p = 2[1 - \Phi(\pm|Z|)] \tag{1.19}$$

Again, we note that the fixed effects model is being used to test whether the studies common effect size is zero or not. This is because the underlying assumption of this model is that all studies share a common effect size. Therefore, when we look at the random effects model, we note that the hypothesis we are testing is different because this model has a different underlying assumption.

Random Effects Model

Unlike the fixed effects model, the random effects model does not assume that the studies included in the meta-analysis share a common true effect size. However, this does not mean that we believe all of the studies give completely different results. We still assume that the studies included are similar otherwise we are violating the general idea of a meta-analysis which is the goal of combining results from similar studies. Therefore, the estimated effect size in this case is a little different. In this

case we also consider the variance between the studies (Borenstein et al., 2009). So, the estimated study effect size is then considered to be

$$Y_i = \mu + \xi_i + \varepsilon_i \tag{1.20}$$

where μ is the mean true effect size, ε_i is the random error for the study, and ξ_i is the variance for the study. Because we consider the variance that occurs between the studies we need to estimate this value before we can calculate the weights and thus the estimated summary effect. So then, we first estimate the between studies variances as

$$T^2 = \frac{Q - df}{C} \tag{1.21}$$

where we calculate Q, df, and C as

$$Q = \sum_{i=1}^{k} W_i Y_i^2 - \frac{\left(\sum_{i=1}^{k} W_i Y_i\right)^2}{\sum_{i=1}^{k} W_i} \tag{1.22}$$

$$df = k - 1 \tag{1.23}$$

$$C = \sum_{i=1}^{k} W_i - \frac{\sum_{i=1}^{k} W_i^2}{\sum_{i=1}^{k} W_i} \tag{1.24}$$

respectively. Note that k is the number of studies that are included in the meta-analysis. We also in this case use the same W_i calculation as the fixed effects model for each study. Also note, this Q that is calculated here will be used later to find heterogeneity. Using this information, we calculate the weights for this model in a different way than the fixed effects model. Under this model the study weights are

$$W_i^* = \frac{1}{V_{Y_i}^*} \tag{1.25}$$

where $V_{Y_i}^*$ is calculated using the estimated between studies variance that was described above as well as the variance of each study.

$$V_{Y_i}^* = V_{Y_i} + T^2 \tag{1.26}$$

Then we use all of this information to calculate the estimated summary effect that accounts for the between studies variance

$$M^* = \frac{\sum_{i=1}^{k} W_i^* Y_i}{\sum_{i=1}^{k} W_i^*} \tag{1.27}$$

The summary effect variance is calculated similar to the fixed effects model, however in this case we use the inverse of the sum of the weights that include the estimate for the between studies variance

$$V_{M^*} = \frac{1}{\sum_{i=1}^{k} W_i^*} \tag{1.28}$$

And then the standard error of the summary effect is again just the square root of the estimated variance for the summary effect

$$SE_{M^*} = \sqrt{V_{M^*}} \tag{1.29}$$

Again, we want to gain a better understanding of the estimated summary effect we found so we calculate a few more values to help. First, the 95% upper and lower confidence limits

$$LL_{M^*} = M^* - 1.96 \times SE_{M^*} \tag{1.30}$$

$$UL_{M^*} = M^* + 1.96 \times SE_{M^*} \tag{1.31}$$

And then we find the Z-value for this model as well however it will be used in a slightly different way than the fixed effects model.

$$Z^* = \frac{M^*}{SE_{M^*}} \tag{1.32}$$

We calculate the one-tailed p-value as

$$p^* = 1 - \Phi(\pm|Z^*|) \tag{1.33}$$

and the two-tailed p-value as

$$p^* = 2[1 - \Phi(\pm|Z^*|)] \tag{1.34}$$

In this case, we are testing the hypothesis that the mean effect size is zero. Therefore, a significant p-value would lead one to reject the null hypothesis and conclude that the mean effect size is not zero. This would mean that using the information from all of the studies could lead a researcher to conclude that there is an effect occurring in the studies.

Differences Between the Models

The main difference between these two methods is that the fixed makes the assumption that the studies of the meta-analysis share a common true effect size and the random does not. Many note that this assumption is very unrealistic and therefore the random effects model is often the better option. Specifically, the fixed effects model considers the variation to only be due to random error whereas the random effects model considers the variation to be due to the random error as well as inherit differences among the studies. Because of this, we see a difference in the interpretation of the summary effect for each of the models. This means that for the fixed effects model the summary effect we calculate is an estimate for the common effect size between the studies because of the assumption that they all share a common effect size. Thus, for the random effects model the summary effect is not an estimate of common effect size because it does not have the same assumption, but rather an estimate of the mean effect size of all of the studies included in the meta-analysis. This difference is apparent when considering the different types of weights that are used in each calculation. Also apparent with consideration of the different weight calculations is that under each model, the influence a study has on the calculation of the summary effect is different. The overall conclusion from this is that when T^2 is non-zero, the weights from each study are more balanced which would be seen in the random effects model. Because of these differences, the null hypothesis that is tested for each of these models is different but easy to understand. For the fixed model, we test the null hypothesis that there is no effect in each study. However, for the random effects model we test the null hypothesis that the mean effect size for all studies is zero. These null hypotheses are consistent with the calculation of the summary effect that is calculated in each model.

Thus, when deciding which of these models to use for a meta-analysis, it is important to consider these differences. To use the fixed effects model, one should be sure that the studies being used meet the assumptions of this model. This means that for the fixed effects model to work

researchers should be sure that the studies share a common effect size, otherwise it will not properly work (Schmidt et al., 2009). It is important to note that when using the fixed effects model, it should be used to estimate the common effect size of the studies used in the meta-analysis, and not to make generalizations that are applied to similar populations (Borenstein et al., 2009). Because of this, the random effects model is often more appropriate to use as the requirements for the fixed effects model are hard to meet. Therefore, the random effects model can be used to try and make conclusions that can be applied to other scenarios.

I.4 Heterogeneity

Heterogeneity Background

The goal of performing a meta-analysis is to gain an understanding on a given topic by using information from the studies conducted for this topic. However, if the studies for this topic produce very different results or effect sizes, it would be wrong to attempt to draw a general conclusion from the summary effect. Thus, an important step in conducting a meta-analysis is to check the heterogeneity. If heterogeneity exists among the studies being used, this means that from study to study the true effect sizes vary (Borenstein et al., 2009). We know that this is an assumption of the random effects model therefore we attempt to gain further understanding on how the studies vary. It is important to note that when we are checking the heterogeneity of the studies, we are only checking the heterogeneity in true effects sizes, and not the effect sizes that we have calculated (Borenstein et al., 2009). We keep this in mind while conducting the test for heterogeneity.

When testing for heterogeneity, we are testing the null hypothesis that the studies of the meta-analysis share a common effect size and the alternative hypothesis would be that the studies do not share a common effect size, or that heterogeneity exists. The first step in testing for heterogeneity is to calculate the Q value for the studies. This value is considered to be the sum of the weighted deviation from the summary effect for each of the studies effect sizes. The calculation is as follows

$$Q = \sum_{i=1}^{k} W_i(Y_i - M)^2 \tag{1.35}$$

where k is the number of studies included in the meta-analysis. Also, W_i is the weight of the studies which is calculated as $1/V_{Y_i}$, or the inverse of the variance of each studies effect size. Thus, Y_i is the effect size that we have calculated for each study that can take any of the forms described in the effect size section. Then M is the summary effect for the studies. We may also calculate Q as

$$Q = \sum_{i=1}^{k} \left(\frac{Y_i - M}{S_i}\right)^2 \tag{1.36}$$

This produces the same value for Q as the previous equation. The S_i is standard error of the study and is just the weight moved inside of the squared term from the previous equation. This form of the equation shows that Q is actually a standardized measure (Borenstein et al., 2009). This is also why we are able to use this equation for any of the forms of effect size. One more form for calculating Q is then

$$Q = \sum_{i=1}^{k} W_i Y_i^2 - \frac{\left(\sum_{i=1}^{k} W_i Y_i\right)^2}{\sum_{i=1}^{k} W_i} \tag{1.37}$$

Each of these formulas will produce the same Q value.

Testing for Heterogeneity

Once we have calculated the Q value, we need the degrees of freedom, df, to calculate the p-value for our calculated Q value. The df for the Q value is one less than the number of studies as seen in equation 23.

Now with both Q and the df, we can calculate the p-value to see whether or not heterogeneity exists between studies. Q has a chi-squared distribution as it is the sum of squared standard normal terms. The level of significance (α) depends mostly on the field for which the meta-analysis is being conducted however it is usually set to either 0.10 or 0.05. A significant p-value is evidence that heterogeneity between studies exists. It is very important to note that a non-significant p-value should

not be taken to mean that there is no heterogeneity, or that the studies share a common effect size (Borenstein et al., 2009). Thus, there are values that are related to Q that can be used to get a better understanding on the variation between the effect sizes in the study.

Understanding Heterogeneity

One of the variables that are used to help understand the differences in effect size between studies is τ^2. τ^2 is the variance of the true effect sizes, however we do not actually know the true effect sizes, so it is estimated using T^2 which is the estimated between studies variance (Borenstein et al., 2009). T^2 is calculated as seen in equation 1.21 and where we calculate C as seen in equation 1.24.

By using $Q - df$, the numerator now represents the variation that is due to differences in the true effect sizes from study to study. Also, C is used to convert the value into the original metric, that was used to find effect size, and then make it an average (Borenstein et al., 2009). T^2 could end up less than zero if $Q < df$ and in this case we just set T^2 equal to zero. In the other case that T^2 is not zero, it will be based on the size of excess variation, the numerator, and will be interpreted un terms of the original metric. Now that we know how T^2 is calculated and what it means, its' use in the random effects model makes sense. Because the random effects model uses the assumption that the true effect sizes vary from study to study, it makes sense that the weights in this model are not only based on the within study variance, V_Y, but also T^2.

Another measure that is used is the standard deviation of the true effect sizes, or τ. Similar to the variance in the previous paragraph, we estimate τ using T. We calculate T as follows

$$T = \sqrt{T^2} \tag{1.38}$$

where T^2 is exactly as previously calculated. Also similar to T^2, T is the same metric as the effect sizes that are calculated for each study. T is used to understand the distribution of the effect sizes for the studies around the summary effect and can give us an understanding of the range of these effect sizes

(Borenstein et al., 2009). This means that even if our test for heterogeneity gives inconclusive results, then we can still attempt to understand how the studies differ from each other.

A final measure that can be used to understand how the studies included in a meta-analysis vary from one another is I^2. This measure was proposed to help understand what portion of the variance is due to actual differences in effect sizes, and not just random error (Higgins et al., 2003). The I^2 ratio can be computed as follows

$$I^2 = \left(\frac{Q - df}{Q}\right) \times 100\% \tag{1.39}$$

which is defined as the ratio of the variation between the studies to the total variation (Borenstein et al., 2009). It can also be written as

$$I^2 = \left(\frac{Varaince_{between}}{Varaince_{total}}\right) \times 100\% \tag{1.40}$$

and

$$I^2 = \left(\frac{\tau^2}{\tau^2 + V_Y}\right) \times 100\% \tag{1.41}$$

When this measure was proposed, it was suggested that the value be considered either low, medium, or high with 25% being low, 50% being medium, and 75% being high (Borenstein et al., 2009). However, those who proposed this idea do note that using this categorization method for all studies is not a good idea and those using it need to consider their specific situation (Higgins et al., 2003). That being said, the Cochrane Reviews does now include the I^2 for their meta-analyses.

Misuse of Heterogeneity Results

Based on the previous sections, it can be seen that testing for heterogeneity and interpreting the results is a very important part of the meta-analysis and can be difficult. Testing for heterogeneity alone should not be used as an indicator of whether the studies all share a common effect size or not. Other measures and interpretations of each studies effect size should be considered after one obtains a non-significant Q value. If the test for heterogeneity leads to the conclusion that there are significant

differences between the studies, then a researcher should not attempt to draw a general conclusion from these studies. This is something that happens often with researchers unfamiliar with the importance of testing for heterogeneity. The I^2 value was proposed with the goal of making it simpler to understand heterogeneity for those unfamiliar with it. However, similar to Q itself, I^2 has been criticized for its low power and inability to meaningfully identify heterogeneity (Baker et al., 2009). Therefore, in the next section we propose a new statistic to test for heterogeneity that is more intuitive for those unfamiliar with advanced statistics.

CHAPTER II: PROPOSING AN ALTERNATIVE HETEROGENEITY TEST

II.1 U-Statistics Background

The type of statistic that we propose using as an alternative to Cochrane's Q is called a U-statistic. To get a better understanding of why we propose this, we give some background on what U-statistics are. By definition, a U-statistic is the average of a symmetric function calculated for the m arguments for a random sample of size n, note that n $\geq$ m (Chen, 2014). Then we have an equation of the form

$$U = \binom{n}{m}^{-1} \sum_{(n,m)} h(X_{i1}, \dots, X_{im}) \tag{2.1}$$

where h is called the kernel with degree m. Specifically, U is an unbiased estimator of a functional of degree m that is defined on a set of distribution functions (Lee, 2019). These statistics are called U-statistics because of their unbiasedness and were named this by Hoeffding who began the study of this class of statistics (Hoeffding, 1992). Note that the U-statistic defined above and by equation 2.1 is referred to as a univariate, one sample U-statistic (Kowalski & Tu, 2008). This is not the only type of U-statistic and there are many more complicated forms that U-statistics can take, however this is the type we propose in this paper. U-statistics are often used as test statistics, for example consider the Mann-Whitney U-Test which is a popular test.

In our case, we apply the concept of U-statistics as follows with regard to equation 2.1. We consider n to the number of studies which we call k in our meta-analysis. In this case, m is set to 2 because we only compare two studies at once. In the next section, we propose a few different U-statistics and the part that varies in each would thus be the kernel function, h. However, they each are symmetric as required for the kernel of the U-statistic. We selected the use of U-statistics for a few reasons. First, we needed to find an alternative to Cochrane's Q because it has been noted to have low statistical power in cases with smaller sample sizes which is often the case in the use of meta-analysis

where the sample size is the number of studies included (Haidich, 2010). Some recommend tests like the Mann-Whitney U-Test in scenarios with small sample sizes (Morgan, 2017, King & Eckersley, 2019). Therefore, we extend the idea of the U-statistic being more beneficial for small sample sizes to the test for heterogeneity. Another reason for proposing U-statistics is the interpretability of those that we propose. Meta-analysis is a widely used statistical tool in a variety of fields and therefore many who use it may not have a statistics background. Therefore, using a test statistic that is more intuitive in its' use in identifying differences in studies may decrease the misuse of meta-analysis with heterogeneous studies. These statistics are considered measures of similarity which is one of the reasons that they are more intuitive. Also, similarity measures are beneficial because they increase from a given minimum (Wolda, 1981). For example, the absolute value of the difference of two numbers would have a minimum of 0 which is very simple to understand. In the next section, we propose a few types of U-statistics.

II.2 Proposed U-Statistics and General Forms

When originally deciding to use U-Statistics, there were three forms that were considered. These forms meet the requirements of being U-statistics such that they are symmetric and are averaged over the number of observations (Chen, 2014). We denote the potential statistics as U_1, U_2, and U_3 such that U_1 is the difference of the absolute value of each term, U_2 is the product of each of the terms, and U_3 is the squared difference of the terms. The terms are defined as the effect size minus the summary effect all over the standard error for each study included in the meta-analysis. This is the same term that is found in the original Cochrane's Q, however for Q the sum of each term squared is used. The U-statistics we used as well as Q are as follows:

$$U_1 = \binom{k}{2}^{-1} \sum_{1 \le i \le k} \left| \left(\frac{Y_i - M}{S_i} \right) - \left(\frac{Y_j - M}{S_j} \right) \right| \tag{2.2}$$

$$U_2 = \binom{k}{2}^{-1} \sum_{1 \le i \le k} \left(\frac{Y_i - M}{S_i} \right) \left(\frac{Y_j - M}{S_j} \right) \tag{2.3}$$

$$U_3 = \binom{k}{2}^{-1} \sum_{1 \le i \le k} \left[\left(\frac{Y_i - M}{S_i} \right) - \left(\frac{Y_j - M}{S_j} \right) \right]^2 \tag{2.4}$$

$$Q = \sum_{1 \le i \le k} \left(\frac{Y_i - M}{S_i} \right)^2 \tag{2.5}$$

To gain a further understanding of these potential statistics, we considered their general form as well as that of the Q-statistics that is normally used to check heterogeneity. First we consider the general form of Q such that we have

$$Q = \sum_{1 \le i \le k} \left(\frac{Y_i - M}{S_i} \right)^2 = \left(\frac{Y_1 - M}{S_1} \right)^2 + \left(\frac{Y_2 - M}{S_2} \right)^2 + \cdots + \left(\frac{Y_k - M}{S_k} \right)^2$$

Next, we consider the general form of U_1 and we have

$$U_1 = \binom{k}{2}^{-1} \sum_{1 \le i \le k} \left| \left(\frac{Y_i - M}{S_i} \right) - \left(\frac{Y_j - M}{S_j} \right) \right|$$

$$= \binom{k}{2}^{-1} \left[\left| \left(\frac{Y_1 - M}{S_1} \right) - \left(\frac{Y_2 - M}{S_2} \right) \right| + \left| \left(\frac{Y_1 - M}{S_1} \right) - \left(\frac{Y_3 - M}{S_3} \right) \right| + \cdots \right.$$

$$+ \left| \left(\frac{Y_1 - M}{S_1} \right) - \left(\frac{Y_k - M}{S_k} \right) \right| + \left| \left(\frac{Y_2 - M}{S_2} \right) - \left(\frac{Y_3 - M}{S_3} \right) \right| + \left| \left(\frac{Y_2 - M}{S_2} \right) - \left(\frac{Y_4 - M}{S_4} \right) \right| + \cdots$$

$$\left. + \left| \left(\frac{Y_2 - M}{S_2} \right) - \left(\frac{Y_k - M}{S_k} \right) \right| + \cdots + \left| \left(\frac{Y_{k-1} - M}{S_{k-1}} \right) - \left(\frac{Y_k - M}{S_k} \right) \right| \right]$$

Then we consider the general form of U_2 which is as follows

$$U_2 = \binom{k}{2}^{-1} \sum_{1 \le i \le k} \left(\frac{Y_i - M}{S_i} \right) \left(\frac{Y_j - M}{S_j} \right)$$

$$= \binom{k}{2}^{-1} \left[\left(\frac{Y_1 - M}{S_1} \right) \left(\frac{Y_2 - M}{S_2} \right) + \left(\frac{Y_1 - M}{S_1} \right) \left(\frac{Y_3 - M}{S_3} \right) + \cdots + \left(\frac{Y_1 - M}{S_1} \right) \left(\frac{Y_k - M}{S_k} \right) \right.$$

$$+ \left(\frac{Y_2 - M}{S_2} \right) \left(\frac{Y_3 - M}{S_3} \right) + \left(\frac{Y_2 - M}{S_2} \right) \left(\frac{Y_4 - M}{S_4} \right) + \cdots + \left(\frac{Y_2 - M}{S_2} \right) \left(\frac{Y_k - M}{S_k} \right) + \cdots$$

$$\left. + \left(\frac{Y_{k-1} - M}{S_{k-1}} \right) \left(\frac{Y_k - M}{S_k} \right) \right]$$

And finally, the general form of U_3 is as follows

$$U_3 = \binom{k}{2}^{-1} \sum_{1 \le i \le k} \left[\left(\frac{Y_i - M}{S_i} \right) - \left(\frac{Y_j - M}{S_j} \right) \right]^2$$

$$= \binom{k}{2}^{-1} \left[\left[\left(\frac{Y_1 - M}{S_1} \right) - \left(\frac{Y_2 - M}{S_2} \right) \right]^2 + \left[\left(\frac{Y_1 - M}{S_1} \right) - \left(\frac{Y_3 - M}{S_3} \right) \right]^2 + \cdots + \left[\left(\frac{Y_1 - M}{S_1} \right) - \left(\frac{Y_k - M}{S_k} \right) \right]^2 \right.$$

$$+ \left[\left(\frac{Y_2 - M}{S_2} \right) - \left(\frac{Y_3 - M}{S_3} \right) \right]^2 + \left[\left(\frac{Y_2 - M}{S_2} \right) - \left(\frac{Y_4 - M}{S_4} \right) \right]^2 + \cdots$$

$$\left. + \left[\left(\frac{Y_2 - M}{S_2} \right) - \left(\frac{Y_k - M}{S_k} \right) \right]^2 + \cdots + \left[\left(\frac{Y_{k-1} - M}{S_{k-1}} \right) - \left(\frac{Y_k - M}{S_k} \right) \right]^2 \right]$$

$$= \frac{2}{k} \left[\left(\frac{Y_1 - M}{S_1} \right)^2 + \left(\frac{Y_2 - M}{S_2} \right)^2 + \cdots + \left(\frac{Y_k - M}{S_k} \right)^2 \right]$$

$$- 2 \binom{k}{2}^{-1} \left[\left(\frac{Y_1 - M}{S_1} \right) \left(\frac{Y_2 - M}{S_2} \right) + \left(\frac{Y_1 - M}{S_1} \right) \left(\frac{Y_3 - M}{S_3} \right) + \cdots + \left(\frac{Y_1 - M}{S_1} \right) \left(\frac{Y_k - M}{S_k} \right) \right.$$

$$+ \left(\frac{Y_2 - M}{S_2} \right) \left(\frac{Y_3 - M}{S_3} \right) + \left(\frac{Y_2 - M}{S_2} \right) \left(\frac{Y_4 - M}{S_4} \right) + \cdots + \left(\frac{Y_2 - M}{S_2} \right) \left(\frac{Y_k - M}{S_k} \right) + \cdots$$

$$\left. + \left(\frac{Y_{k-1} - M}{S_{k-1}} \right) \left(\frac{Y_k - M}{S_k} \right) \right]$$

Each of these proposed U-statistics is symmetric and allows us to measure the differences between the studies included in the meta-analysis.

II.3 Relationship Between U-Statistics and Cochrane's Q

From the general forms seen above, it could be concluded that using U_3 may be unnecessary. If we further inspect U_3, it can be observed that it is actually of the form

$$U_3 = \frac{2}{k} Q - 2\binom{k}{2}^{-1} U_2 \qquad (2.6)$$

We see that the U-statistic U_3 actually contains the Q-statistic and therefore may not tell us anything new. Thus, we make the decision to not further explore the use of U_3 moving forward. From here, we

can get a better understanding of how Q and U_2 are related by considering the sum of the standardized effect sizes. We know that the sum of standardized effect sizes will equal 0 such that we have

$$\frac{Y_1 - M}{S_1} + \frac{Y_2 - M}{S_2} + \cdots + \frac{Y_k - M}{S_k} = 0 \tag{2.7}$$

And we also know that if we square both sides of equation 2.7 we have

$$\left(\frac{Y_1 - M}{S_1} + \frac{Y_2 - M}{S_2} + \cdots + \frac{Y_k - M}{S_k}\right)^2 = 0 \tag{2.8}$$

We can then expand equation 2.8 such and we get the following

$$\begin{gathered}\left(\frac{Y_1 - M}{S_1}\right)^2 + \left(\frac{Y_2 - M}{S_2}\right)^2 + \cdots + \left(\frac{Y_k - M}{S_k}\right)^2 \\ +2\left[\begin{array}{c}\left(\frac{Y_1 - M}{S_1}\right)\left(\frac{Y_2 - M}{S_2}\right) + \left(\frac{Y_1 - M}{S_1}\right)\left(\frac{Y_3 - M}{S_3}\right) + \cdots \\ + \left(\frac{Y_1 - M}{S_1}\right)\left(\frac{Y_k - M}{S_k}\right) + \left(\frac{Y_2 - M}{S_2}\right)\left(\frac{Y_3 - M}{S_3}\right) + \left(\frac{Y_2 - M}{S_2}\right)\left(\frac{Y_4 - M}{S_4}\right) + \cdots \\ + \left(\frac{Y_2 - M}{S_2}\right)\left(\frac{Y_k - M}{S_k}\right) + \cdots + \left(\frac{Y_{k-1} - M}{S_{k-1}}\right)\left(\frac{Y_k - M}{S_k}\right)\end{array}\right] = 0\end{gathered} \tag{2.9}$$

By considering the general forms of Q and U_2 from above we can conclude that equation 2.9 can be rewritten as

$$Q + 2\binom{k}{2}U_2 = 0 \tag{2.10}$$

Therefore, we can conclude that actually Q and U_2 are related to each other in some form. However, we continue looking into the use of U_2 because it is still different from Q due to the averaging nature and symmetry. Obviously, we also continue to consider U_1 moving forward as its general form does not appear to pose any issues and it is the most inuitive form. We keep these general forms and conclusions drawn from them in mind as we continue on in this paper. Moving forward we refer to U_1 as the absolute value U-statistic and we refer to U_2 as the product U-statistic.

CHAPTER III: APPLICATIONS OF PROPOSED U-STATISTICS

III.1 Using Simulation to Approximate the Distribution

Because we want to use the U-statistics that have been proposed to test heterogeneity, we must be able to calculate a p-value. To do so, we use a simulation method in R to simulate values for the U-statistic for meta-analysis with studies ranging from three to fifteen. This method is often referred to as the Monte Carlo method which consists of repeating a simulation to use to test for statistical significance (Zintzaras & Ioannidis, 2005). Because of the way that the U-statistics we proposed are designed, they include a standardized value made up of the study effect size minus the summary effect all over the standard error of the study. Therefore, it is easy to have R create random standardized values to simulate the values that produce the U-statistics. For each U-statistic, we simulate scenarios in which k, the number of studies, is 3 through 10 and then also 15, however it would be easy to extend this method to scenarios with more studies included. The R code for these simulations is found in the appendix.

To begin, we consider the absolute value U-statistic and how we simulate it. First, for each scenario, as stated previously, we randomize a standard normal value to represent the standardized value obtained for each study. Then we let U equal the sum of the absolute value of the difference between each pair of random standard normal values all over the total number of comparisons. Because these values are randomized, we can replicate the U-statistics many times to gain an understanding on how the absolute value U-statistic is distributed. Using R, we replicate this process of simulating U a million times. We create a histogram for each scenario that can be seen in the *Figure 1*. From these histograms, we can see that even if the number of studies changes, the values for this absolute value U-statistic is always distributed around about 1. Also, we note that the values for the absolute value U-statistic cannot be less than 0, so this will be a one-tailed test. Therefore, when we calculate the p-value we will consider what percentage of simulated values are more extreme, or larger, than our calculated

absolute value U-statistic. This method of finding the p-value will be used in the next section when we apply the new methods for testing heterogeneity. In *Table 1*, we give the critical values for varying significance levels and numbers of studies. We then repeat this for the product U-statistic.

Distribution Simulations for Absolute Value U-Statistic

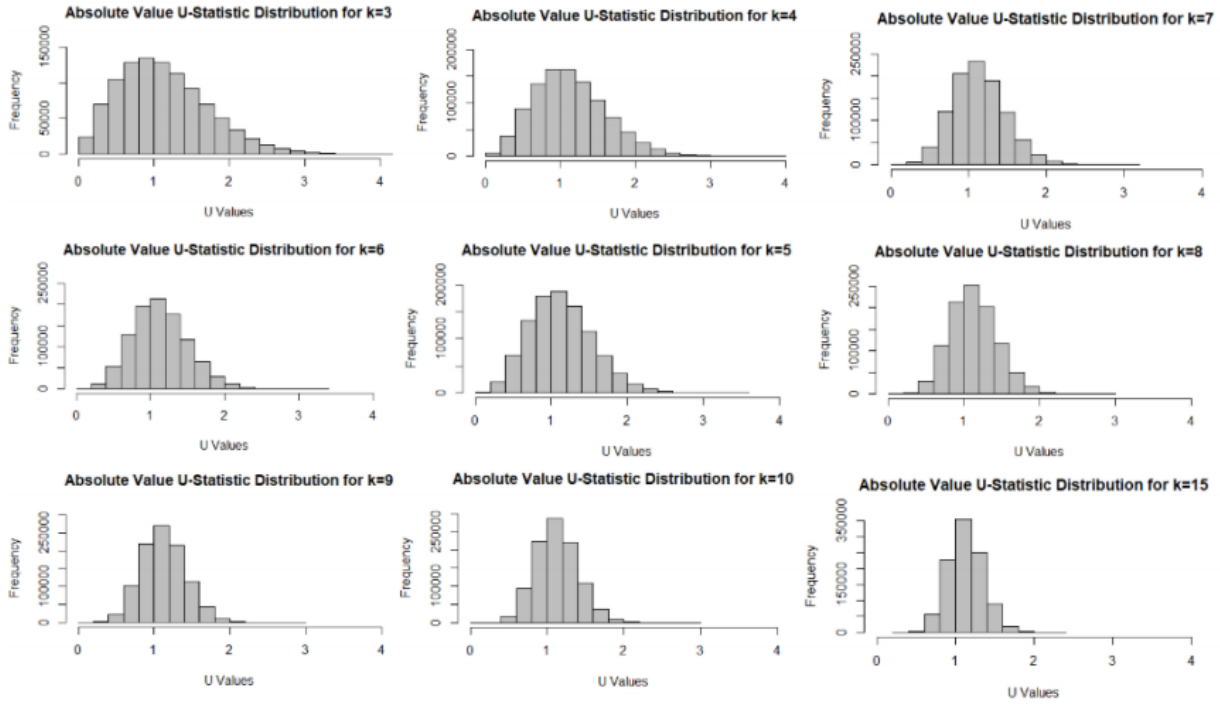

Figure 1 This figure shows the simulated values for the absolute value U-statistic for scenarios with three through ten studies and then also fifteen. These simulated values are then constructed into a histogram to represent the distribution for this statistic.

Critical Values Table for Absolute Value U-Statistic

Absolute Value U-Statistic			
Number of Studies	Significance Level		
	1%	5%	10%
k=3	2.7479	2.2083	1.9333
k=4	2.3967	1.9842	1.7707
k=5	2.1995	1.8575	1.6792
k=6	2.0749	1.7724	1.6173
k=7	1.9793	1.7098	1.5705
k=8	1.9076	1.6628	1.5355
k=9	1.8564	1.6272	1.5088
k=10	1.8090	1.5954	1.4853
k=15	1.6648	1.4992	1.4123

Table 1 This table contains the critical values for each of the distributions simulated in *Figure 1*. They are calculated for 1%, 5%, and 10% significance levels. This would only be used as a one-tailed test because the closer to zero the U-statistic, the more similar the studies.

The process for simulating the product U-statistic is exactly the same as that for the absolute value U-statistic except for one difference. Rather than the using the sum of the absolute value of the differences, we use the sum of the product of each pair of studies all over the number of comparisons. From here, we again replicate the U value a million times and create histograms of the values produced. In *Figure 2* we see that the values are distributed around zero for each of the different study sizes. Unlike the test using the absolute value, the product U-statistic will be a two-tailed test because the U values in this case can be negative. Therefore, when calculating the p-value for this statistic, we find the number of studies that are more extreme in the given direction depending on whether the calculated value is positive or negative. Then, we double this number and take is as a percentage of the total million studies because we see that this statistic would be symmetric. We then use this process to calculate the p-value while using this statistic to test for heterogeneity. In *Table 2*, we give the critical values for varying number of studies and significance levels for others to use when testing for heterogeneity.

Distribution Simulations for Product U-Statistic

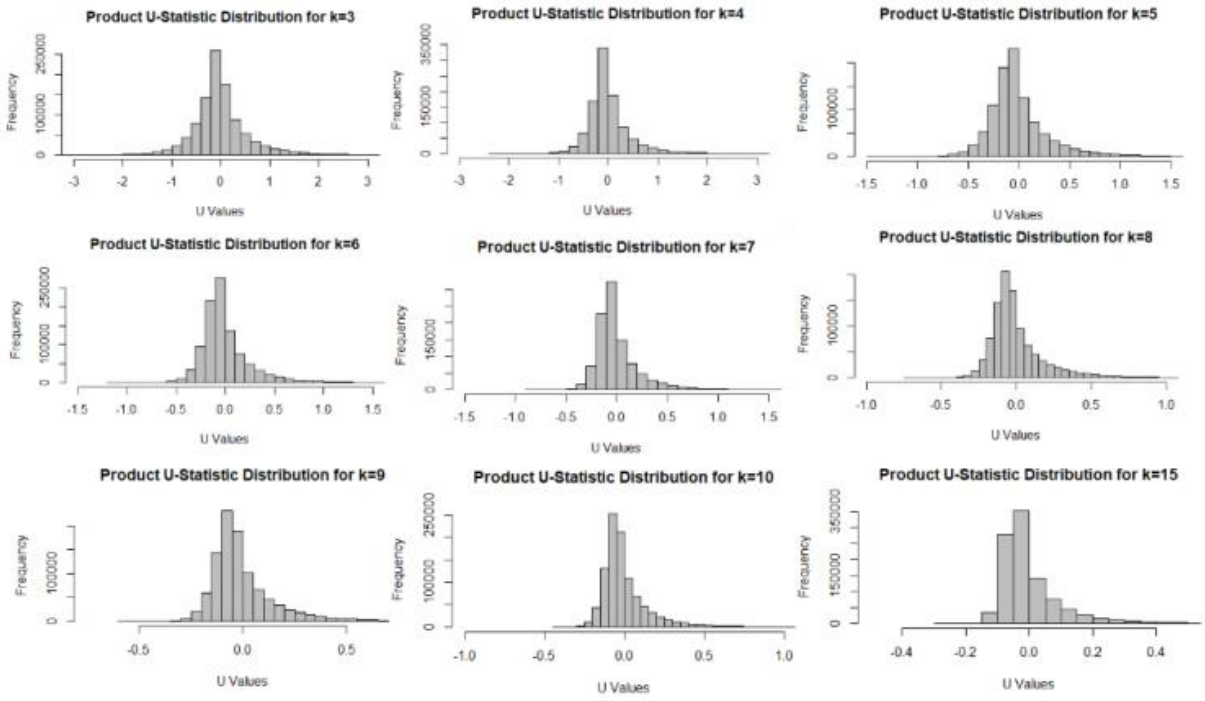

Figure 2 This figure shows the simulated values for the product U-statistic for scenarios with three through ten studies and then also fifteen. These simulated values are then constructed into a histogram to represent the distribution for this statistic just as the previous figure for the absolute value U-statistic.

Critical Values Table for Product U-Statistic

Product U-Statistic			
Number of Studies	Significance Level		
	1%	5%	10%
k=3	1.9511	1.0234	0.6489
k=4	1.4458	0.7473	0.4681
k=5	1.1492	0.5945	0.3707
k=6	0.9562	0.4922	0.3033
k=7	0.8177	0.4190	0.2583
k=8	0.7111	0.3644	0.2235
k=9	0.6299	0.3223	0.1982
k=10	0.5694	0.2887	0.1764
k=15	0.3777	0.1923	0.1167

Table 2 This table contains the critical values for each of the distributions simulated in *Figure 1*. They are calculated for 1%, 5%, and 10% significance levels. Unlike the absolute value U-statistic, this one would be used as a two-tail test.

III.2 An Example of How We Use U-Statistics

Working Through an Example

To fully understand the difference in testing heterogeneity by using a U-statistic rather than the Cochrane's Q-statistic, we work through an example. The data for this example is found in the Introduction to Meta-Analysis Textbook written by Borenstein et al. in 2009. This example includes 6 studies, and we find the summary effect, test heterogeneity, and draw conclusions to fully understand what these statistics give us. To start, we calculate the summary effect of these studies using the random effects model. We chose this model because for these particular studies we do not want to assume that their true effect sizes are equal. To do so, we begin by estimating the effect sizes for each of the studies by using the standardized mean difference that was discussed previously. These estimates are seen in *Table 4* in the effect size column and were calculated using the information listed in *Table 3*. We obtain an estimate for the summary effect, or M^* value, of 0.356, a V_{M^*} of 0.11, and an SE_{M^*} of 0.104. Using these, we calculate the lower and upper limits of the 95% confidence interval for the

summary effect to be 0.152 and 0.560. Because this is just an example we calculate both the one-tail and two-tail p-values using the Z^* value of 3.417 and we get 0.0003 and 0.0006, respectively.

After estimating the summary effect, we then can use this information to test for heterogeneity using Cochrane's Q and the two proposed U-statistics. We can see the results of each heterogeneity test in *Table 5*. Cochrane's Q and the absolute value U-statistic would both lead us to reject the null hypothesis and conclude there is heterogeneity, however the product U-statistic does not. In *Figure 3*, we see the forest plot created for the estimated effect sizes of the studies. This figure can be used to make a visual guess as to whether or not heterogeneity exists between the studies and compared to the results of the heterogeneity tests. However, visual inspection does not provide much clarity. For Cochrane's Q, recall that we often use a few other measures to understand the heterogeneity such as T^2 and I^2. We obtain a T^2 value of 0.036 and an I^2 value of 57.42%. We further analyze these findings in the results section.

Data for Initial Example Studies

Introduction to Meta-Analysis Worked Example						
Study	**Treatment n**	**Treatment mean**	**Treatment SD**	**Control n**	**Control mean**	**Control SD**
Carroll	60	94	22	60	92	20
Grant	65	98	21	65	92	22
Peck	40	98	28	40	88	26
Donat	200	94	19	200	82	17
Stewart	50	98	21	45	88	22
Young	85	96	21	85	92	22

Table 3 This table contains data for all 6 of the studies included in this example for both the treatment and control groups. These values are used to calculate the effect sizes for each of the studies.

Effect Sizes for Initial Example Studies

Study	Effect Size	Variance Within	Weights	
	Y	V_Y	W	W^*
Carroll	0.095	0.033	30.303	14.492
Grant	0.277	0.031	32.258	14.925
Peck	0.367	0.050	20.000	11.628
Donat	0.664	0.011	90.909	21.277
Stewart	0.462	0.043	23.256	12.658
Young	0.185	0.023	43.478	16.949
Sum	-	-	240.204	91.929

Table 4 This table includes the estimated effect sizes for each study, the variance and the weights using the fixed and random effects methods.

Forest Plot of Effect Sizes for Initial Example Studies

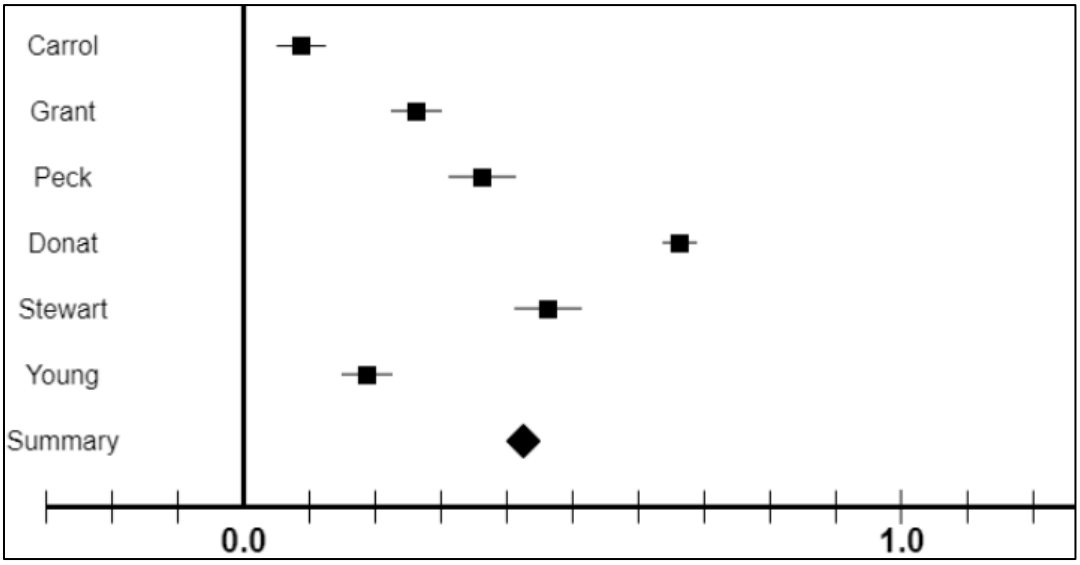

Figure 3 This figure is the forest plot that contains each of the estimated effect sizes for the studies, represented as boxes. The lines coming from the boxes represent the standard error for each estimate. The diamond in the bottom represents the estimated summary effect for these studies.

Heterogeneity Test Results for Initial Example

Test Type	Heterogeneity Test Statistic	P-Value
Cochrane's Q	$Q = 11.743$	$p = 0.0428$
Absolute Value U-Statistic	$U_{AV} = 1.770$	$p = 0.0495$
Product U-Statistic	$U_P = -0.321$	$p = 0.0787$

Table 5 This table is the summary of the three heterogeneity tests used in this example, Cochrane's Q and the two proposed U-statistics.

III.3 Application of Using U-Statistics

Similar to the previous example worked from the Introduction to Meta-Analysis textbook, we fully work through a real world example to gain more understanding on the proposed test statistics. The data for these real world examples were taken from the Cochrane Library which is a database of systematic reviews (Parker & Handoll, 2010). These specific examples involve data that compares two types of hip surgery techniques: gamma nail and sliding hip screw. We work two examples. The first is comparing the length of time in minutes for each of the techniques and the other compares the blood loss in mL for each of the techniques. Many studies have been done to see if the sliding hip screw method is better than the gamma nail method as the gamma nail method has been known to cause complications in patients (Saarenpää et al., 2009).

For both of these scenarios, the random effects model is used. Note that the fixed effects model is rarely appropriate because of its' assumption that the studies true effect sizes are the same. Thus, in each case we start by estimating the summary effect of the studies. After this, we test for heterogeneity between studies using Cochrane's Q and then comparing to the two proposed U-statistics. Finally, we attempt to draw conclusions from these results and gain a further understanding on how the heterogeneity tests differ.

Length of Surgery

We begin by finding the summary effect for the studies conducted on the length of surgery for the gamma nail and sliding hip screw methods. There are 6 studies included and *Table 6* lists the number

of observations, means, and standard deviations for both surgery types in each study. Again, we first find the effect size for these studies by using the standardized mean difference between the gamma nail and sliding hip screw surgeries because we are interested in the difference between these two methods, and we standardize because these are two different methods used. *Table 7* lists the estimated effect sizes for each of the studies, as well as their variance and weights (with the last column specifically being the weights under the random effects model). Using this information, we calculate the estimated summary effect $M^*$to be 0.30, V_{M^*} to be 0.0298, and SE_{M^*} to be 0.173. Using these, we are able to calculate the lower and upper limits of the 95% confidence interval for the summary effect to be -0.039 and 0.639, respectively. We calculate the two-tail p-value using the Z^* value of 1.743 and we get 0.083.

Data for Length of Surgery Studies

Length of Surgery (minutes)						
Study	**GN n**	**GN mean**	**GN SD**	**SHS n**	**SHS mean**	**SHS SD**
1	53	59	23.9	49	47	13.3
2	31	56.7	17	36	54.3	16.4
3	60	47.1	20.8	60	53.4	8.3
4	203	55.4	20	197	61.3	22.2
5	104	46	11	106	44	15
6	73	65	29	73	51	22

Table 6, GN = Gamma Nail, SHS = Sliding Hip Screw

This table contains data for all 6 of the studies included in the length of surgery application for both the GN and SHS groups. These values are used to calculate the effect sizes for each of the studies.

Effect Sizes for Length of Surgery Studies

Study	Effect Size	Variance Within	Weights	
	Y	V_Y	W	W^*
1	-0.279	0.010	100	6.289
2	-0.398	0.034	29.412	5.464
3	0.544	0.026	38.462	5.714
4	0.614	0.037	27.027	5.376
5	0.152	0.019	52.632	5.952
6	0.144	0.060	16.667	4.785
Sum	-	-	264.2	33.580

Table 7 This table includes the estimated effect sizes for each study using the standardized mean difference method, the variance and the weights using the fixed and random effects methods.

After we find the summary effect, we move onto testing for heterogeneity. When considering the forest plot in *Figure 4* for the estimated effect sizes of each study, we could assume that heterogeneity does exist between these studies because of the spread of values observed. However, we check test the heterogeneity using these statistics to verify they work. In *Table 8* we see the results of the three tests considered. All of these tests give extremely significant p-values and lead us to conclude that heterogeneity exists among the studies. We also obtain a T^2 value of 0.149 and an I^2 value of 85.84%. Again, we further analyze these findings in the results section.

Forest Plot of Effect Sizes for Length of Surgery Studies

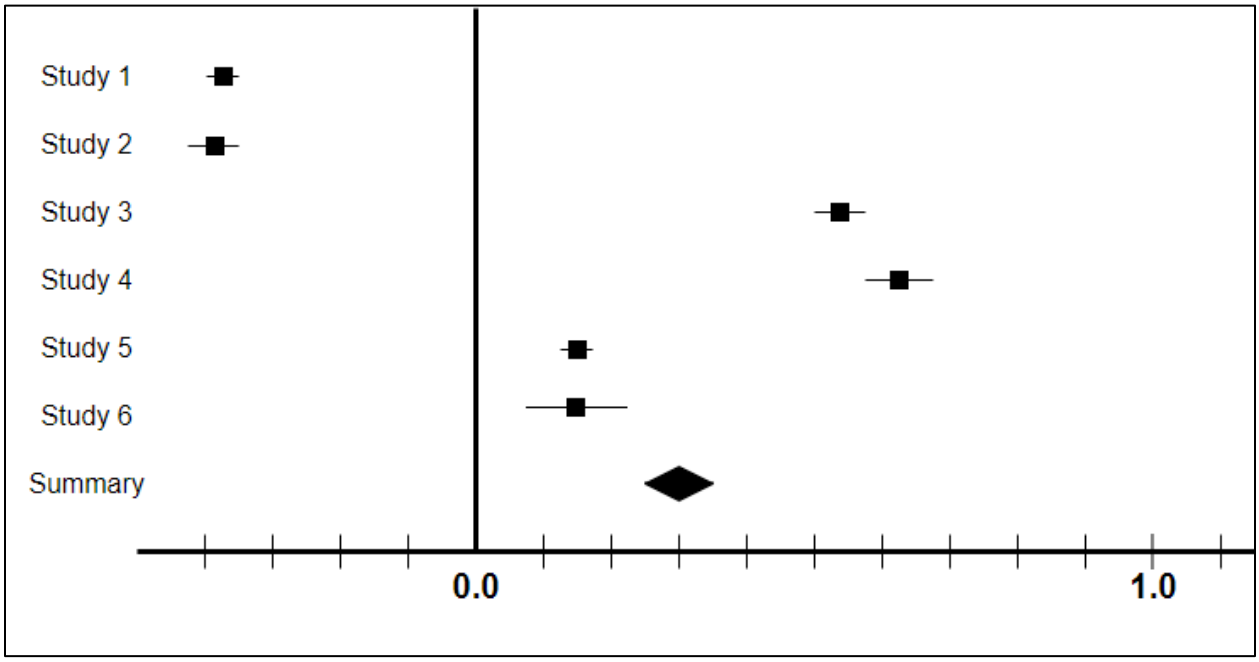

Figure 4 This figure is the forest plot that contains each of the estimated effect sizes for the studies, represented as boxes. The lines coming from the boxes represent the standard error for each estimate. The diamond in the bottom represents the estimated summary effect for these studies.

Heterogeneity Test Results for Length of Surgery Application

Test Type	Heterogeneity Test Statistic	P-Value
Cochrane's Q	$Q = 35.314$	$p < 0.0001$
Absolute Value U-Statistic	$U_{AV} = 3.194$	$p = 0.000001$
Product U-Statistic	$U_P = -1.029$	$p = 0.000018$

Table 8 This table is the summary of the three heterogeneity tests used in the length of surgery application, Cochrane's Q and the two proposed U-statistics.

Finally, we use one more application to compare the proposed tests for checking heterogeneity among studies. In this application, we consider the blood loss in mL in 5 studies comparing the gamma nail and sliding hip screw surgery methods as before. In *Table 9*, the number of observations, the means, and the standard deviations are listed for both the GN and SHS methods. This information is then used to calculate the estimates of the effect sizes that are seen in *Table 10*. These were calculated using the

standardized mean difference for the same reasons as it was used for the length of surgery scenario. Again, we estimate the summary effect $M^*$to be -0.144 with a 95% confidence interval lower limit of -0.299 and upper limit of 0.011. We get this interval using a V_{M^*} of 0.006 and a SE_{M^*} of 0.079. We calculate the two-tail p-value using the Z^* value of -1.82 and we get 0.069.

Data for Blood Loss Studies

Blood Loss (mL)						
Study	**GN n**	**GN mean**	**GN SD**	**SHS n**	**SHS mean**	**SHS SD**
1	93	814	548	93	1043	508
2	52	258.7	145.4	49	259.2	137.5
3	60	152.3	130.7	60	160.3	110.8
4	203	244.4	384.9	197	260.4	325.5
5	73	240	190	73	280	280

Table 9, GN = Gamma Nail, SHS = Sliding Hip Screw

This table contains data for all 5 of the studies included in the blood loss application for both the GN and SHS groups. These values are used to calculate the effect sizes for each of the studies.

Effect Sizes for Blood Loss Studies

Study	**Effect Size** Y	**Variance Within** V_Y	**Weights** W	W^*
1	-0.433	0.022	45.455	33.003
2	-0.004	0.040	25.228	20.704
3	-0.066	0.033	29.984	24.213
4	-0.045	0.010	99.952	54.645
5	-0.167	0.028	36.373	27.548
Sum	-	-	236.992	160.113

Table 10 This table includes the estimated effect sizes for each study using the standardized mean difference method, the variance and the weights using the fixed and random effects methods

After finding the summary effect, we test for heterogeneity. We consider the forest plot in *Figure 5* for the estimated effect sizes of each study, and we note that most studies seem similar however one effect size is larger than the others. We then test for heterogeneity. In *Table* 11, we see the results of the three tests considered. All of these tests give non-significant p-values and lead us to conclude that heterogeneity does not exist among the studies despite one differing studies effect size.

We also obtain a T^2 value of 0.008 and an I^2 value of 26.61%. We analyze these findings in the results section.

Forest Plot of Effect Sizes for Blood Loss Studies

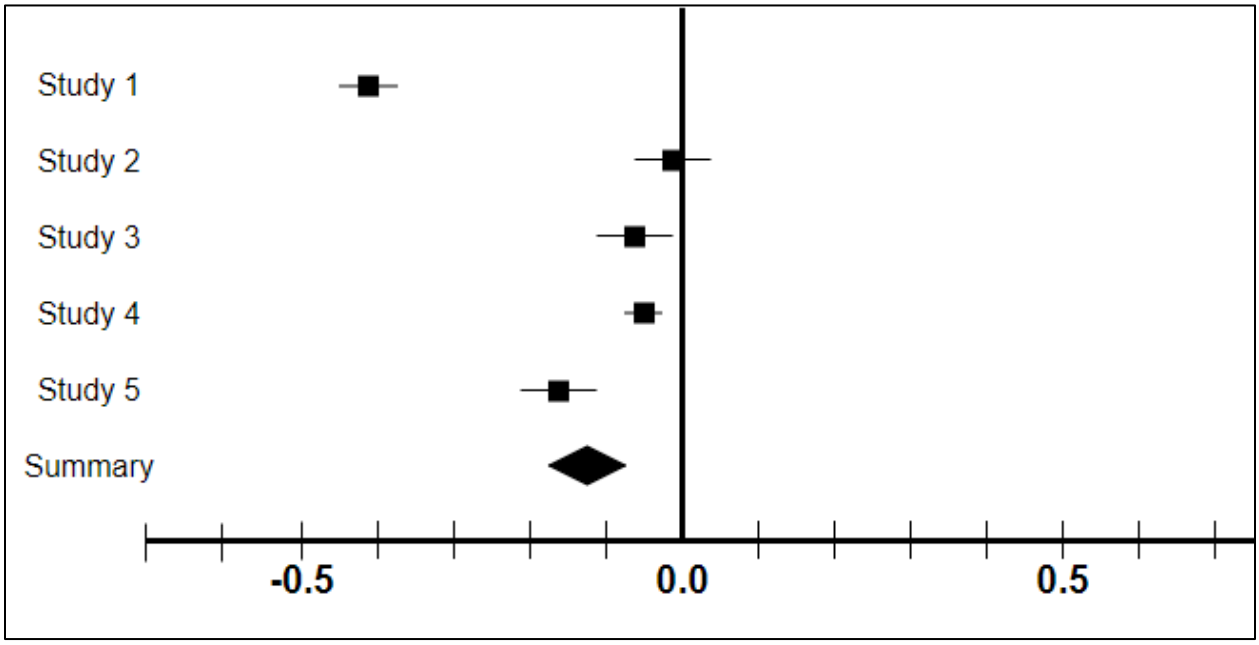

Figure 5 This figure is the forest plot that contains each of the estimated effect sizes for the studies, represented as boxes. The lines coming from the boxes represent the standard error for each estimate. The diamond in the bottom represents the estimated summary effect for these studies.

Heterogeneity Test Results for Blood Loss Application

Test Type	Heterogeneity Test Statistic	P-Value
Cochrane's Q	$Q = 5.4560$	$p = 0.2436$
Absolute Value U-Statistic	$U_{AV} = 1.338$	$p = 0.2913$
Product U-Statistic	$U_P = -0.271$	$p = 0.2447$

Table 11 This table is the summary of the three heterogeneity tests used in the blood loss application, Cochrane's Q and the two proposed U-statistics.

CHAPTER IV: RESULTS

IV.1 Interpretation of Example

We begin by interpreting the results of our initial example, specifically we begin by interpreting the estimated summary effect. Again, using the random effects model we estimated a summary effect of 0.356 (0.152, 0.560). For this estimated summary effect, we test the null hypothesis that the mean effect size is zero. For this test we got a two-tailed p-value of 0.0006 which would lead us to reject the null hypothesis. This means that for whatever treatment was implemented for these studies, their estimated mean effect was greater than zero. Knowing this, we move on to the results of the tests for heterogeneity.

The results of each test for heterogeneity, seen in *Table 5*, give us slightly different results. We see that our test using Cochrane's Q gives us a Q value of 11.743 with a p-value of 0.0428. This would lead us to reject the null hypothesis and conclude that there heterogeneity does exist between the studies. When we look at the absolute value U-statistic, we see different results. We get a U-value of 1.770 with a p-value of 0.0495 which would still lead us to reject the null hypothesis, but in this case it is much closer to a non-significant difference. Finally, we see the most different results when we calculate the product U-statistic. We get a U-value of -0.321 with a p-value of 0.0787 that leads us to fail to reject the null hypothesis and conclude that there is not a significant difference in the studies. These finding are all consistent with the interpretation of the forest plot in *Figure 3* which were inconclusive based on a visual inspection. Many note that when results are this close to non-significant, conclusions should not be drawn (Borenstein et al., 2009). Note, we also obtained a T^2 of 0.036 which is the estimate of the variance between the true effect sizes. Also, we have an I^2 of 57.42% which is the percent of variance that represent real differences in effect size. Some would consider this a moderate amount but recall that this is a somewhat bias measure and generally has low power.

IV.2 Interpretation of Applications

In this section, we move onto the interpretation of the results for the real world application of the proposed method. For each scenario, we again begin with the interpretation of the estimated summary effect and then move onto the comparison for the heterogeneity tests.

Length of Surgery Results

First, we look at the results for the length of surgery scenario in which we compare the gamma nail and sliding hip screw methods. We estimate a summary effect of 0.30 (-0.039, 0.639). Again, because we used the random effects model, we are testing the null hypothesis that the mean effect size is zero. Here, we calculate the two-tailed test p-value to be 0.083 so we fail to reject the null hypothesis. In other words, this means that the on average there does not appear to be a significant difference in the length of the surgery between the gamma nail and sliding hip screw measures. However, before we can make this conclusion, we must test for heterogeneity to see if the studies are even similar enough to make this conclusion.

The results for each of the three tests for heterogeneity are seen in *Table 8*. We observe that each of the tests give a similar result in this case. For Cochrane's Q we get a Q value of 35.314 and a p-value of about 0.0001 which leads us to reject the null hypothesis and conclude that heterogeneity does exist between the studies included in the meta-analysis. The absolute value U-statistic is calculated as 3.194 and gives us a p-value of 0.000001 which gives us the same conclusion as the previous test. Finally, we get the same result from the product U-statistic which is calculated as -1.029 with a p-value of 0.000018. This is supported by the forest plot for this scenario seen in *Figure 4* which upon visual inspection would lead to a conclusion that the studies are very different. If one would want to consider the I^2 of 85.84% it would also appear to support this conclusion as it is relatively high. Therefore, we cannot draw a general conclusion about the difference in length of surgery between the two methods using information from these given studies as they are too different.

Blood Loss Results

Finally, we interpret the results of the studies comparing the blood loss between the two hip surgery methods. We estimate the summary effect to be -0.144 (-0.299, 0.011). When testing the null hypothesis in this case we get a two-tailed p-value of 0.069 so we fail to reject the null hypothesis and conclude that the mean, or summary, effect size does not differ from zero. We then check the heterogeneity of these studies to see if we can draw a conclusion for these results.

In *Table 11*, we see the heterogeneity test results for each of the three methods considered. Again, in this application we see similar results among each of the three tests. For Cochrane's Q we get a Q value of 5.4560 with a p-value of 0.2436, therefore we would fail to reject the null hypothesis and conclude that there is not a significant difference between the studies. For the absolute value U-statistic, we come to the same conclusion with a U value of 1.338 and a p-value of 0.02913. Similarly, we get a product U-statistic of -0.271 and a p-value of 0.2447 which leads us to conclude there is not significant difference between studies. The forest plot in *Figure 5* also appears to show little difference, except for one observation that appears different from the others. We also calculate an I^2 value of 26.61% which would appear to be relatively low, although we should not use this as our only reasoning. All of these results could mean that these methods are not sensitive to outliers, however this may not cause an issue.

CHAPTER V: DISCUSSION

V.1 Comparing Cochrane's Q and the U-Statistics

When we consider the U-statistics that were proposed, one of their benefits is that their use is more intuitive than the Cochrane's Q. For example, the absolute value U-statistic is really just the difference between the two standardized effect sizes of a pair of studies. For somebody that is not very familiar with statistical methods, using the absolute value U-statistic test would naturally make the most sense. There are a few more obvious differences between Cochrane's Q and the U-statistics. We know that Q is impacted much more by outliers than the U-statistics, especially for small sample sizes. This is due to the fact that the U-statistics are averaged over the number of comparisons used. Averaging has the effect of minimizing the impact of an outlier. Other than that, the methods proposed do give comparable results to those produced using Cochrane's Q. Another important point to note is that the use of meta-analysis is expanding into many different fields. For example, meta-analysis is becoming more commonly used with Life Cycle Analyses (LCA) which contain a lot of information (Brandao et al., 2012). In this case, a simpler approach may be easier to implement. Therefore, the U-statistics could be a more approachable alternative for those in non-statistical fields that wish to fully understand the meta-analysis that they are implementing.

V.2 Limitations

There were a few limitations in proposing these methods as alternatives to Cochrane's Q. For one, time was a major limitation. Realistically, we would want to perform many simulations to compare the proposed U-statistics against Cochrane's Q. However, this process is lengthy and was not able to be accomplished in the given amount of time. Another limitation is the lack of knowledge in regard to the theory of U-statistics. Although this paper does cover some of the basic details of U-statistics and what they are used for, a much deeper understanding of them would be beneficial. One reason this would be beneficial is that it would allow us to better understand the distribution of these variables in a way that

the simulation of the U-statistics does not allow. Overcoming these two limitations would greatly strengthen the proposal of these as alternatives to Cochrane's Q. The benefit of overcoming these limitations is further explained in the next section.

V.3 Future Research Directions

As stated previously, we would have ideally conducted a simulation study to further understand the use of these U-statistics in comparison to Cochrane's Q. Therefore, in future research, data should be simulated to use for conducting heterogeneity tests using the proposed statistics. One of the benefits of a simulation study is that we would then be able to compare the efficiency of the proposed U-statistics to that of Cochrane's Q. A benefit of using these U-statistics is that they can be used as a general framework to move forward with. These similarity measures could be modified in meaningful ways to better fit certain situations as long as they still meet the requirements of U-statistics. For example, these statistics could be improved upon by adding meaningful weights to them based on the sample sizes of the studies (Ciol et al., 2006). Also, these proposed U-statistics could be modified in a way that makes them more useful for testing in specific fields. This would require a deeper understanding of U-statistics and the forms of U-statistics that are most appropriate for specific scenarios and data types. Therefore, future research should also focus on deeper application of U-statistic theory, as noted in the limitations. Although there are a few areas where these methods could be improved, these U-statistics appear to be a promising and realistic alternative to using Cochrane's Q to test for heterogeneity among the studies of a meta-analysis.

REFERENCES

[1] Baker, W. L., Michael White, C., Cappelleri, J. C., Kluger, J., Coleman, C. I., & From the Health Outcomes, Policy, and Economics (HOPE) Collaborative Group. (2009). Understanding heterogeneity in meta-analysis: the role of meta-regression. International journal of clinical practice, 63(10), 1426-1434.

[2] Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). Introduction to meta-analysis. John Wiley & Sons.

[3] Brandao, M., Heath, G., & Cooper, J. (2012). What can meta-analyses tell us about the reliability of life cycle assessment for decision support?.

[4] Chen, X. (2014). U-Statistics. Wiley StatsRef: Statistics Reference Online, 1-6.

[5] Ciol, M. A., Hoffman, J. M., Dudgeon, B. J., Shumway-Cook, A., Yorkston, K. M., & Chan, L. (2006). Understanding the use of weights in the analysis of data from multistage surveys. Archives of physical medicine and rehabilitation, 87(2), 299-303.

[6] Fletcher, J. (2007). What is heterogeneity and is it important?. Bmj, 334(7584), 94-96.

[7] Gorard, S. (2005). Revisiting a 90-year-old debate: the advantages of the mean deviation. British Journal of Educational Studies, 53(4), 417-430.

[8] Haidich, A. B. (2010). Meta-analysis in medical research. Hippokratia, 14(Suppl 1), 29.

[9] Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in medicine, 21(11), 1539-1558.

[10] Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. Bmj, 327(7414), 557-560.

[11] Hoeffding, W. (1992). A class of statistics with asymptotically normal distribution. In Breakthroughs in statistics (pp. 308-334). Springer, New York, NY.

[12] Jones, D. R. (1995). Meta-analysis: weighing the evidence. Statistics in medicine, 14(2), 137-149.

[13] Jones, J. B., Blecker, S., & Shah, N. R. (2008). Meta-analysis 101: What you want to know in the era of comparative effectiveness. American health & drug benefits, 1(3), 38.

[14] King, A. P., & Eckersley, R. J. (2019). Inferential statistics III: nonparametric hypothesis testing. Statistics for Biomedical Engineers and Scientists, AP King and RJ Eckersley, Eds, 119-145.

[15] Kolasa, J., & Rollo, C. D. (1991). Introduction: the heterogeneity of heterogeneity: a glossary. In Ecological heterogeneity (pp. 1-23). Springer, New York, NY.

[16] Kowalski, J., & Tu, X. M. (2008). Modern applied U-statistics (Vol. 714). John Wiley & Sons.

[17] Lee, A. J. (2019). U-statistics: Theory and Practice. Routledge.

[18] Ma, Y., & Mazumdar, M. (2011). Multivariate meta-analysis: a robust approach based on the theory of U-statistic. Statistics in Medicine, 30(24), 2911-2929.

[19] Morgan, C. J. (2017). Use of proper statistical techniques for research studies with small samples. American Journal of Physiology-Lung Cellular and Molecular Physiology, 313(5), L873-L877.

[20] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. Journal of Clinical Epidemiology, 134, 103-112.

[21] Parker, M. J., & Handoll, H. H. (2010). Gamma and other cephalocondylic intramedullary nails versus extramedullary implants for extracapsular hip fractures in adults. Cochrane database of systematic reviews, (9).

[22] Saarenpää, I., Heikkinen, T., Ristiniemi, J., Hyvönen, P., Leppilahti, J., & Jalovaara, P. (2009). Functional comparison of the dynamic hip screw and the Gamma locking nail in trochanteric hip fractures: a matched-pair study of 268 patients. International orthopaedics, 33(1), 255-260.

[23] Schmidt, F. L., Oh, I. S., & Hayes, T. L. (2009). Fixed-versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. British Journal of Mathematical and Statistical Psychology, 62(1), 97-128.

[24] Wolda, H. (1981). Similarity indices, sample size and diversity. Oecologia, 50(3), 296-302.

[25] Zintzaras, E., & Ioannidis, J. P. (2005). Heterogeneity testing in meta-analysis of genome searches. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 28(2), 123-137.

APPENDIX: R CODE

The R code used to simulate the distributions for both the proposed absolute value U-statistic and the product U-statistic are included in the appendix.

ABSOLUTE VALUE U-STATISTIC R CODE

```
###distribution for U-stat where n=3###
U_distr3 = function(x_1,x_2,x_3){
 Ustat= (1/3)*(abs(x_1-x_2)+abs(x_1-x_3)
              +abs(x_2-x_3))
 return(Ustat)
}
distribution3 = replicate(1000000, U_distr3(x_1=rnorm(1,mean=0,sd=1),
                                  x_2=rnorm(1,mean=0,sd=1),
                                  x_3=rnorm(1,mean=0,sd=1)))
distribution3
hist(distribution3)
###distribution for U-stat where n=4###
U_distr4 = function(x_1,x_2,x_3,x_4){
 Ustat= (1/6)*(abs(x_1-x_2)+abs(x_1-x_3)+abs(x_1-x_4)+
              +abs(x_2-x_3)+abs(x_2-x_4)
             +abs(x_3-x_4))

 return(Ustat)
}
distribution4 = replicate(1000000, U_distr4(x_1=rnorm(1,mean=0,sd=1),
```

```
                                x_2=rnorm(1,mean=0,sd=1),

                                x_3=rnorm(1,mean=0,sd=1),

                                x_4=rnorm(1,mean=0,sd=1)))
distribution4

hist(distribution4)

###distribution for U-stat where n=5###

U_distr5 = function(x_1,x_2,x_3,x_4,x_5){

  Ustat= (1/10)*(abs(x_1-x_2)+abs(x_1-x_3)+abs(x_1-x_4)+abs(x_1-x_5)+

                +abs(x_2-x_3)+abs(x_2-x_4)+abs(x_2-x_5)

               +abs(x_3-x_4)+abs(x_3-x_5)+abs(x_4-x_5))

  return(Ustat)

}

distribution5 = replicate(1000000, U_distr5(x_1=rnorm(1,mean=0,sd=1),

                                x_2=rnorm(1,mean=0,sd=1),

                                x_3=rnorm(1,mean=0,sd=1),

                                x_4=rnorm(1,mean=0,sd=1),

                                x_5=rnorm(1,mean=0,sd=1)))

distribution5

hist(distribution5)


###distribution for U-stat where n=6###

U_distr6 = function(x_1,x_2,x_3,x_4,x_5,x_6){

  Ustat= (1/15)*(abs(x_1-x_2)+abs(x_1-x_3)+abs(x_1-x_4)+abs(x_1-x_5)+

                abs(x_1-x_6)+abs(x_2-x_3)+abs(x_2-x_4)+abs(x_2-x_5)+abs(x_2-x_6)
```

```
        +abs(x_3-x_4)+abs(x_3-x_5)+abs(x_3-x_6)+abs(x_4-x_5)+abs(x_4-x_6)+abs(x_5-x_6))
 return(Ustat)
}
distribution6 = replicate(1000000, U_distr6(x_1=rnorm(1,mean=0,sd=1),
                                            x_2=rnorm(1,mean=0,sd=1),
                                            x_3=rnorm(1,mean=0,sd=1),
                                            x_4=rnorm(1,mean=0,sd=1),
                                            x_5=rnorm(1,mean=0,sd=1),
                                            x_6=rnorm(1,mean=0,sd=1)))
distribution6
hist(distribution6)
###distribution for U-stat where n=7###
U_distr7 = function(x_1,x_2,x_3,x_4,x_5,x_6, x_7){
 Ustat= (1/21)*(abs(x_1-x_2)+abs(x_1-x_3)+abs(x_1-x_4)+abs(x_1-x_5)+
          abs(x_1-x_6)+abs(x_1-x_7)+abs(x_2-x_3)+abs(x_2-x_4)+abs(x_2-x_5)+abs(x_2-
x_6)+abs(x_2-x_7)
        +abs(x_3-x_4)+abs(x_3-x_5)+abs(x_3-x_6)+abs(x_3-x_7)+abs(x_4-x_5)+abs(x_4-
x_6)+abs(x_4-x_7)+abs(x_5-x_6)+abs(x_5-x_7)
        +abs(x_6-x_7) )
 return(Ustat)
}
distribution7 = replicate(1000000, U_distr7(x_1=rnorm(1,mean=0,sd=1),
                                            x_2=rnorm(1,mean=0,sd=1),
                                            x_3=rnorm(1,mean=0,sd=1),
```

```
                                x_4=rnorm(1,mean=0,sd=1),

                                x_5=rnorm(1,mean=0,sd=1),

                                x_6=rnorm(1,mean=0,sd=1),

                                x_7=rnorm(1,mean=0,sd=1)))

distribution7

hist(distribution7)

###distribution for U-stat where n=8###

U_distr8 = function(x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8){

  Ustat= (1/28)*(abs(x_1-x_2)+abs(x_1-x_3)+abs(x_1-x_4)+abs(x_1-x_5)+

                   abs(x_1-x_6)+abs(x_1-x_7)+abs(x_1-x_8)

                 +abs(x_2-x_3)+abs(x_2-x_4)+abs(x_2-x_5)+abs(x_2-x_6)+abs(x_2-x_7)+abs(x_2-x_8)

                 +abs(x_3-x_4)+abs(x_3-x_5)+abs(x_3-x_6)+abs(x_3-x_7)+abs(x_3-x_8)

                 +abs(x_4-x_5)+abs(x_4-x_6)+abs(x_4-x_7)+abs(x_4-x_8)

                 +abs(x_5-x_6)+abs(x_5-x_7)+abs(x_5-x_8)

                 +abs(x_6-x_7)+abs(x_6-x_8)

                 +abs(x_7-x_8))

  return(Ustat)

}

distribution8 = replicate(1000000, U_distr8(x_1=rnorm(1,mean=0,sd=1),

                                x_2=rnorm(1,mean=0,sd=1),

                                x_3=rnorm(1,mean=0,sd=1),

                                x_4=rnorm(1,mean=0,sd=1),

                                x_5=rnorm(1,mean=0,sd=1),

                                x_6=rnorm(1,mean=0,sd=1),
```

```
                                x_7=rnorm(1,mean=0,sd=1),

                                x_8=rnorm(1,mean=0,sd=1)))

distribution8

hist(distribution8)

###distribution for U-stat where n=9###

U_distr9 = function(x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9){

  Ustat= (1/36)*(abs(x_1-x_2)+abs(x_1-x_3)+abs(x_1-x_4)+abs(x_1-x_5)+

                   abs(x_1-x_6)+abs(x_1-x_7)+abs(x_1-x_8)+abs(x_1-x_9)

                 +abs(x_2-x_3)+abs(x_2-x_4)+abs(x_2-x_5)+abs(x_2-x_6)+abs(x_2-x_7)+abs(x_2-

x_8)+abs(x_2-x_9)

                 +abs(x_3-x_4)+abs(x_3-x_5)+abs(x_3-x_6)+abs(x_3-x_7)+abs(x_3-x_8)+abs(x_3-x_9)

                 +abs(x_4-x_5)+abs(x_4-x_6)+abs(x_4-x_7)+abs(x_4-x_8)+abs(x_4-x_9)

                 +abs(x_5-x_6)+abs(x_5-x_7)+abs(x_5-x_8)+abs(x_5-x_9)

                 +abs(x_6-x_7)+abs(x_6-x_8)+abs(x_6-x_9)

                 +abs(x_7-x_8)+abs(x_7-x_9)

                 +abs(x_8-x_9))

  return(Ustat)

}

distribution9 = replicate(1000000, U_distr9(x_1=rnorm(1,mean=0,sd=1),

                                x_2=rnorm(1,mean=0,sd=1),

                                x_3=rnorm(1,mean=0,sd=1),

                                x_4=rnorm(1,mean=0,sd=1),

                                x_5=rnorm(1,mean=0,sd=1),

                                x_6=rnorm(1,mean=0,sd=1),
```

```
                                x_7=rnorm(1,mean=0,sd=1),

                                x_8=rnorm(1,mean=0,sd=1),

                                x_9=rnorm(1,mean=0,sd=1)))

distribution9

hist(distribution9)

###distribution for U-stat where n=10###

U_distr10 = function(x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_10){

  Ustat= (1/45)*(abs(x_1-x_2)+abs(x_1-x_3)+abs(x_1-x_4)+abs(x_1-x_5)+

                   abs(x_1-x_6)+abs(x_1-x_7)+abs(x_1-x_8)+abs(x_1-x_9)+abs(x_1-x_10)

                 +abs(x_2-x_3)+abs(x_2-x_4)+abs(x_2-x_5)+abs(x_2-x_6)+abs(x_2-x_7)+abs(x_2-

x_8)+abs(x_2-x_9)+abs(x_2-x_10)

                 +abs(x_3-x_4)+abs(x_3-x_5)+abs(x_3-x_6)+abs(x_3-x_7)+abs(x_3-x_8)+abs(x_3-

x_9)+abs(x_3-x_10)

                 +abs(x_4-x_5)+abs(x_4-x_6)+abs(x_4-x_7)+abs(x_4-x_8)+abs(x_4-x_9)+abs(x_4-x_10)

                 +abs(x_5-x_6)+abs(x_5-x_7)+abs(x_5-x_8)+abs(x_5-x_9)+abs(x_5-x_10)

                 +abs(x_6-x_7)+abs(x_6-x_8)+abs(x_6-x_9)+abs(x_6-x_10)

                 +abs(x_7-x_8)+abs(x_7-x_9)+abs(x_7-x_10)

                 +abs(x_8-x_9)+abs(x_8-x_10)

                 +abs(x_9-x_10)

                 )

  return(Ustat)

}

distribution10 = replicate(1000000, U_distr10(x_1=rnorm(1,mean=0,sd=1),

                                x_2=rnorm(1,mean=0,sd=1),
```

```
                    x_3=rnorm(1,mean=0,sd=1),
                    x_4=rnorm(1,mean=0,sd=1),
                    x_5=rnorm(1,mean=0,sd=1),
                    x_6=rnorm(1,mean=0,sd=1),
                    x_7=rnorm(1,mean=0,sd=1),
                    x_8=rnorm(1,mean=0,sd=1),
                    x_9=rnorm(1,mean=0,sd=1),
                    x_10=rnorm(1,mean=0,sd=1)))
distribution10
hist(distribution10)
###distribution for U-stat where n=15###
U_distr15 = function(x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_10,x_11,x_12,x_13,x_14,x_15){
  Ustat= (1/105)*(abs(x_1-x_2)+abs(x_1-x_3)+abs(x_1-x_4)+abs(x_1-x_5)+
            abs(x_1-x_6)+abs(x_1-x_7)+abs(x_1-x_8)+abs(x_1-x_9)+abs(x_1-x_10)+abs(x_1-
x_11)+abs(x_1-x_12)+abs(x_1-x_13)+abs(x_1-x_14)+abs(x_1-x_15)
          +abs(x_2-x_3)+abs(x_2-x_4)+abs(x_2-x_5)+abs(x_2-x_6)+abs(x_2-x_7)+abs(x_2-
x_8)+abs(x_2-x_9)+abs(x_2-x_10)+abs(x_2-x_11)+abs(x_2-x_12)+abs(x_2-x_13)+abs(x_2-
x_14)+abs(x_2-x_15)
          +abs(x_3-x_4)+abs(x_3-x_5)+abs(x_3-x_6)+abs(x_3-x_7)+abs(x_3-x_8)+abs(x_3-
x_9)+abs(x_3-x_10)+abs(x_3-x_11)+abs(x_3-x_12)+abs(x_3-x_13)+abs(x_3-x_14)+abs(x_3-x_15)
          +abs(x_4-x_5)+abs(x_4-x_6)+abs(x_4-x_7)+abs(x_4-x_8)+abs(x_4-x_9)+abs(x_4-
x_10)+abs(x_4-x_11)+abs(x_4-x_12)+abs(x_4-x_13)+abs(x_4-x_14)+abs(x_4-x_15)
          +abs(x_5-x_6)+abs(x_5-x_7)+abs(x_5-x_8)+abs(x_5-x_9)+abs(x_5-x_10)+abs(x_5-
x_11)+abs(x_5-x_12)+abs(x_5-x_13)+abs(x_5-x_14)+abs(x_5-x_15)
```

```
        +abs(x_6-x_7)+abs(x_6-x_8)+abs(x_6-x_9)+abs(x_6-x_10)+abs(x_6-x_11)+abs(x_6-
x_12)+abs(x_6-x_13)+abs(x_6-x_14)+abs(x_6-x_15)
        +abs(x_7-x_8)+abs(x_7-x_9)+abs(x_7-x_10)+abs(x_7-x_11)+abs(x_7-x_12)+abs(x_7-
x_13)+abs(x_7-x_14)+abs(x_7-x_15)
        +abs(x_8-x_9)+abs(x_8-x_10)+abs(x_8-x_11)+abs(x_8-x_12)+abs(x_8-x_13)+abs(x_8-
x_14)+abs(x_8-x_15)
        +abs(x_9-x_10)+abs(x_9-x_11)+abs(x_9-x_12)+abs(x_9-x_13)+abs(x_9-x_14)+abs(x_9-
x_15)
        +abs(x_10-x_11)+abs(x_10-x_12)+abs(x_10-x_13)+abs(x_10-x_14)+abs(x_10-x_15)
        +abs(x_11-x_12)+abs(x_11-x_13)+abs(x_11-x_14)+abs(x_11-x_15)
        +abs(x_12-x_13)+abs(x_12-x_14)+abs(x_12-x_15)
        +abs(x_13-x_14) +abs(x_13-x_15)
        +abs(x_14-x_15)
 )
 return(Ustat)
}
distribution15 = replicate(1000000, U_distr15(x_1=rnorm(1,mean=0,sd=1),
                                x_2=rnorm(1,mean=0,sd=1),
                                x_3=rnorm(1,mean=0,sd=1),
                                x_4=rnorm(1,mean=0,sd=1),
                                x_5=rnorm(1,mean=0,sd=1),
                                x_6=rnorm(1,mean=0,sd=1),
                                x_7=rnorm(1,mean=0,sd=1),
                                x_8=rnorm(1,mean=0,sd=1),
```

```
                                    x_9=rnorm(1,mean=0,sd=1),
                                    x_10=rnorm(1,mean=0,sd=1),
                                    x_11=rnorm(1,mean=0,sd=1),
                                    x_12=rnorm(1,mean=0,sd=1),
                                    x_13=rnorm(1,mean=0,sd=1),
                                    x_14=rnorm(1,mean=0,sd=1),
                                    x_15=rnorm(1,mean=0,sd=1)))
        distribution15
        hist(distribution15)
```

PRODUCT U-STATISTIC R CODE

```
        ###distribution for U-stat where n=3###
        U_distr3 = function(x_1,x_2,x_3){
          Ustat= (1/3)*((x_1*x_2)+(x_1*x_3)
                      +(x_2*x_3))
          return(Ustat)
        }
        distribution3 = replicate(1000000, U_distr3(x_1=rnorm(1,mean=0,sd=1),
                                    x_2=rnorm(1,mean=0,sd=1),
                                    x_3=rnorm(1,mean=0,sd=1)))
        distribution3
        hist(distribution3)
        ###distribution for U-stat where n=4###
        U_distr4 = function(x_1,x_2,x_3,x_4){
          Ustat= (1/6)*((x_1*x_2)+(x_1*x_3)+(x_1*x_4)+
```

```
              +(x_2*x_3)+(x_2*x_4)
             +(x_3*x_4))
  return(Ustat)
}
distribution4 = replicate(1000000, U_distr4(x_1=rnorm(1,mean=0,sd=1),
                                  x_2=rnorm(1,mean=0,sd=1),
                                  x_3=rnorm(1,mean=0,sd=1),
                                  x_4=rnorm(1,mean=0,sd=1)))
distribution4
hist(distribution4)
###distribution for U-stat where n=5###
U_distr5 = function(x_1,x_2,x_3,x_4,x_5){
  Ustat= (1/10)*((x_1*x_2)+(x_1*x_3)+(x_1*x_4)+(x_1*x_5)+
               +(x_2*x_3)+(x_2*x_4)+(x_2*x_5)
             +(x_3*x_4)+(x_3*x_5)+(x_4*x_5))
  return(Ustat)
}
distribution5 = replicate(1000000, U_distr5(x_1=rnorm(1,mean=0,sd=1),
                                  x_2=rnorm(1,mean=0,sd=1),
                                  x_3=rnorm(1,mean=0,sd=1),
                                  x_4=rnorm(1,mean=0,sd=1),
                                  x_5=rnorm(1,mean=0,sd=1)))
distribution5
hist(distribution5)
```

```
###distribution for U-stat where n=6###
U_distr6 = function(x_1,x_2,x_3,x_4,x_5,x_6){
  Ustat= (1/15)*((x_1*x_2)+(x_1*x_3)+(x_1*x_4)+(x_1*x_5)+
            (x_1*x_6)+(x_2*x_3)+(x_2*x_4)+(x_2*x_5)+(x_2*x_6)
            +(x_3*x_4)+(x_3*x_5)+(x_3*x_6)+(x_4*x_5)+(x_4*x_6)+(x_5*x_6))
  return(Ustat)
}
distribution6 = replicate(1000000, U_distr6(x_1=rnorm(1,mean=0,sd=1),
                                x_2=rnorm(1,mean=0,sd=1),
                                x_3=rnorm(1,mean=0,sd=1),
                                x_4=rnorm(1,mean=0,sd=1),
                                x_5=rnorm(1,mean=0,sd=1),
                                x_6=rnorm(1,mean=0,sd=1)))
hist(distribution6)
###distribution for U-stat where n=7###
U_distr7 = function(x_1,x_2,x_3,x_4,x_5,x_6,x_7){
  Ustat= (1/21)*((x_1*x_2)+(x_1*x_3)+(x_1*x_4)+(x_1*x_5)+
              (x_1*x_6)+(x_1*x_7)
            +(x_2*x_3)+(x_2*x_4)+(x_2*x_5)+(x_2*x_6)+(x_2*x_7)
            +(x_3*x_4)+(x_3*x_5)+(x_3*x_6)+(x_3*x_7)
            +(x_4*x_5)+(x_4*x_6)+(x_4*x_7)
            +(x_5*x_6)+(x_5*x_7)
            +(x_6*x_7))
  return(Ustat)
```

```
}

distribution7 = replicate(1000000, U_distr7(x_1=rnorm(1,mean=0,sd=1),
                                  x_2=rnorm(1,mean=0,sd=1),
                                  x_3=rnorm(1,mean=0,sd=1),
                                  x_4=rnorm(1,mean=0,sd=1),
                                  x_5=rnorm(1,mean=0,sd=1),
                                  x_6=rnorm(1,mean=0,sd=1),
                                  x_7=rnorm(1,mean=0,sd=1)))

distribution7

hist(distribution7)

###distribution for U-stat where n=8###

U_distr8 = function(x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8){
  Ustat= (1/28)*((x_1*x_2)+(x_1*x_3)+(x_1*x_4)+(x_1*x_5)+
                   (x_1*x_6)+(x_1*x_7)+(x_1*x_8)
                 +(x_2*x_3)+(x_2*x_4)+(x_2*x_5)+(x_2*x_6)+(x_2*x_7)+(x_2*x_8)
                 +(x_3*x_4)+(x_3*x_5)+(x_3*x_6)+(x_3*x_7)+(x_3*x_8)
                 +(x_4*x_5)+(x_4*x_6)+(x_4*x_7)+(x_4*x_8)
                 +(x_5*x_6)+(x_5*x_7)+(x_5*x_8)
                 +(x_6*x_7)+(x_6*x_8)
                 +(x_7*x_8))
  return(Ustat)
}

distribution8 = replicate(1000000, U_distr8(x_1=rnorm(1,mean=0,sd=1),
                                  x_2=rnorm(1,mean=0,sd=1),
```

```
                                x_3=rnorm(1,mean=0,sd=1),

                                x_4=rnorm(1,mean=0,sd=1),

                                x_5=rnorm(1,mean=0,sd=1),

                                x_6=rnorm(1,mean=0,sd=1),

                                x_7=rnorm(1,mean=0,sd=1),

                                x_8=rnorm(1,mean=0,sd=1)))
distribution8

hist(distribution8)

###distribution for U-stat where n=9###

U_distr9 = function(x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9){

  Ustat= (1/36)*((x_1*x_2)+(x_1*x_3)+(x_1*x_4)+(x_1*x_5)+

                   (x_1*x_6)+(x_1*x_7)+(x_1*x_8)+(x_1*x_9)

                 +(x_2*x_3)+(x_2*x_4)+(x_2*x_5)+(x_2*x_6)+(x_2*x_7)+(x_2*x_8)+(x_2*x_9)

                 +(x_3*x_4)+(x_3*x_5)+(x_3*x_6)+(x_3*x_7)+(x_3*x_8)+(x_3*x_9)

                 +(x_4*x_5)+(x_4*x_6)+(x_4*x_7)+(x_4*x_8)+(x_4*x_9)

                 +(x_5*x_6)+(x_5*x_7)+(x_5*x_8)+(x_5*x_9)

                 +(x_6*x_7)+(x_6*x_8)+(x_6*x_9)

                 +(x_7*x_8)+(x_7*x_9)

                 +(x_8*x_9))

  return(Ustat)

}

distribution9 = replicate(1000000, U_distr9(x_1=rnorm(1,mean=0,sd=1),

                                x_2=rnorm(1,mean=0,sd=1),

                                x_3=rnorm(1,mean=0,sd=1),
```

```
                                    x_4=rnorm(1,mean=0,sd=1),

                                    x_5=rnorm(1,mean=0,sd=1),

                                    x_6=rnorm(1,mean=0,sd=1),

                                    x_7=rnorm(1,mean=0,sd=1),

                                    x_8=rnorm(1,mean=0,sd=1),

                                    x_9=rnorm(1,mean=0,sd=1)))

distribution9

hist(distribution9)

###distribution for U-stat where n=10###

U_distr10 = function(x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_10){

  Ustat= (1/45)*((x_1*x_2)+(x_1*x_3)+(x_1*x_4)+(x_1*x_5)+

            (x_1*x_6)+(x_1*x_7)+(x_1*x_8)+(x_1*x_9)+(x_1*x_10)

+(x_2*x_3)+(x_2*x_4)+(x_2*x_5)+(x_2*x_6)+(x_2*x_7)+(x_2*x_8)+(x_2*x_9)+(x_2*x_10)

            +(x_3*x_4)+(x_3*x_5)+(x_3*x_6)+(x_3*x_7)+(x_3*x_8)+(x_3*x_9)+(x_3*x_10)

            +(x_4*x_5)+(x_4*x_6)+(x_4*x_7)+(x_4*x_8)+(x_4*x_9)+(x_4*x_10)

            +(x_5*x_6)+(x_5*x_7)+(x_5*x_8)+(x_5*x_9)+(x_5*x_10)

            +(x_6*x_7)+(x_6*x_8)+(x_6*x_9)+(x_6*x_10)

            +(x_7*x_8)+(x_7*x_9)+(x_7*x_10)

            +(x_8*x_9)+(x_8*x_10)

            +(x_9*x_10))

  return(Ustat)

}

distribution10 = replicate(1000000, U_distr10(x_1=rnorm(1,mean=0,sd=1),

                                    x_2=rnorm(1,mean=0,sd=1),
```

```
                    x_3=rnorm(1,mean=0,sd=1),
                    x_4=rnorm(1,mean=0,sd=1),
                    x_5=rnorm(1,mean=0,sd=1),
                    x_6=rnorm(1,mean=0,sd=1),
                    x_7=rnorm(1,mean=0,sd=1),
                    x_8=rnorm(1,mean=0,sd=1),
                    x_9=rnorm(1,mean=0,sd=1),
                    x_10=rnorm(1,mean=0,sd=1)))
distribution10
hist(distribution10)
###distribution for U-stat where n=15###
U_distr15 = function(x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_10,x_11,x_12,x_13,x_14,x_15){
  Ustat= (1/105)*((x_1*x_2)+(x_1*x_3)+(x_1*x_4)+(x_1*x_5)+
(x_1*x_6)+(x_1*x_7)+(x_1*x_8)+(x_1*x_9)+(x_1*x_10)+(x_1*x_11)+(x_1*x_12)+(x_1*x_13)+(x_
1*x_14)+(x_1*x_15)
+(x_2*x_3)+(x_2*x_4)+(x_2*x_5)+(x_2*x_6)+(x_2*x_7)+(x_2*x_8)+(x_2*x_9)+(x_2*x_10)+(x_2*
x_11)+(x_2*x_12)+(x_2*x_13)+(x_2*x_14)+(x_2*x_15)
+(x_3*x_4)+(x_3*x_5)+(x_3*x_6)+(x_3*x_7)+(x_3*x_8)+(x_3*x_9)+(x_3*x_10)+(x_3*x_11)+(x_3
*x_12)+(x_3*x_13)+(x_3*x_14)+(x_3*x_15)
+(x_4*x_5)+(x_4*x_6)+(x_4*x_7)+(x_4*x_8)+(x_4*x_9)+(x_4*x_10)+(x_4*x_11)+(x_4*x_12)+(x_
4*x_13)+(x_4*x_14)+(x_4*x_15)
+(x_5*x_6)+(x_5*x_7)+(x_5*x_8)+(x_5*x_9)+(x_5*x_10)+(x_5*x_11)+(x_5*x_12)+(x_5*x_13)+(x
_5*x_14)+(x_5*x_15)
```

```
+(x_6*x_7)+(x_6*x_8)+(x_6*x_9)+(x_6*x_10)+(x_6*x_11)+(x_6*x_12)+(x_6*x_13)+(x_6*x_14)+(
x_6*x_15)
+(x_7*x_8)+(x_7*x_9)+(x_7*x_10)+(x_7*x_11)+(x_7*x_12)+(x_7*x_13)+(x_7*x_14)+(x_7*x_15)
            +(x_8*x_9)+(x_8*x_10)+(x_8*x_11)+(x_8*x_12)+(x_8*x_13)+(x_8*x_14)+(x_8*x_15)
            +(x_9*x_10)+(x_9*x_11)+(x_9*x_12)+(x_9*x_13)+(x_9*x_14)+(x_9*x_15)
            +(x_10*x_11)+(x_10*x_12)+(x_10*x_13)+(x_10*x_14)+(x_10*x_15)
            +(x_11*x_12)+(x_11*x_13)+(x_11*x_14)+(x_11*x_15)
            +(x_12*x_13)+(x_12*x_14)+(x_12*x_15)
            +(x_13*x_14)+(x_13*x_15)
            +(x_14*x_15))
  return(Ustat)
}
distribution15 = replicate(1000000, U_distr15(x_1=rnorm(1,mean=0,sd=1),
                                    x_2=rnorm(1,mean=0,sd=1),
                                    x_3=rnorm(1,mean=0,sd=1),
                                    x_4=rnorm(1,mean=0,sd=1),
                                    x_5=rnorm(1,mean=0,sd=1),
                                    x_6=rnorm(1,mean=0,sd=1),
                                    x_7=rnorm(1,mean=0,sd=1),
                                    x_8=rnorm(1,mean=0,sd=1),
                                    x_9=rnorm(1,mean=0,sd=1),
                                    x_10=rnorm(1,mean=0,sd=1),
                                    x_11=rnorm(1,mean=0,sd=1),
                                    x_12=rnorm(1,mean=0,sd=1),
```

```
                    x_13=rnorm(1,mean=0,sd=1),
                    x_14=rnorm(1,mean=0,sd=1),
                    x_15=rnorm(1,mean=0,sd=1)))
distribution15
hist(distribution15)
```