AMSI VACATION RESEARCH SCHOLARSHIPS 2019–20

EXPLORE THE MATHEMATICAL SCIENCES THIS SUMMER

Effect Size Measures: Sampling behaviour and use across fields of study

Mikayla Goodwin
Supervised by Dr Robert King
University of Newcastle

Vacation Research Scholarships are funded jointly by the Department of Education and the Australian Mathematical Sciences Institute.

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Australian Government Department of Education

1 Acknowledgements

I am grateful for assistance provided by Virginia Walker in providing guidance and advice in researching the use of statistics in other fields of study.

I would also like to acknowledge the guidance provided by Dr Robert King in preparing and supervising this research project.

2 Introduction

In statistics, most studies are interested in whether there are detectable differences between various variables or models, whether it be the difference in means between two groups or the significance that a particular variable has in determining the outcome in a linear model. Often in statistics, there is a large focus on the hypothesis test. While this is a great measure to explore if there is a relationship, it often overlooks the magnitude or direction of this relationship. Therefore, various insights from experiments and data are often overlooked. Effect sizes however are constructed in order to analyse and conceptualise the relationships occurring in the data. They are able to provide a clearer understanding of the interaction between variables and groups. Beyond finding if there exists a significant effect, the effect size allows to measure whether that effect is important.

Just as there are many different forms of data, many statistical tests and many different insights people are interested in finding from their data, there also exists a large number of effect size measures. Effect size measures each have a different use, different strengths and weaknesses and therefore are utilised across a variety of different contexts. An understanding of effect size allows for a more robust and deeper understanding of one's data, the relationships that exist amongst their data as well as what these relationships are truly representing.

Effect sizes are useful to report on, as due to the structure of many statistical tests, there is potential for manipulation of results, whether knowing or unknowingly. When thinking of data analysis, many would consider a larger number of samples to be a positive. However, due to many statistical tests being biased towards larger samples, meaning larger samples are more likely to return statistically significant results, it can be easy to misrepresent the reality of the relationships in the data. This is where effect size becomes important, as effect sizes aim to measure and conceptualise the relationships that this statistical significance is testing.

This comes from the concept of statistical power, the probability that a hypothesis test rejects the null hypothesis when an alternative hypothesis is true. As the number of samples increases, statistical tests become more likely to detect smaller differences as being statistically significant. When considering that small differences are more noticeable at larger sample sizes, reporting on the effect size would show that the real difference is small. While this could still be an important result in some areas, failing to include it in other areas may be ambiguous.

An implication from here is that whilst statistical significance is highly dependant and sensitive to sample size, effect size measures are supposedly independent of sample sizes. This will be explored throughout a simulation portion of this paper, which aims to capture the long-term behaviour of effect size measures as sample sizes increase. Whether their results are consistent and the performance of various measures throughout a range of different effect sizes and sample sizes.

Another portion of this paper will also be discussing the usage of statistics across various fields of study. As many areas of research include statistical analysis, a literature review will be conducted to observe the language and methods that non-statistician academics use when approaching data analysis. This involved considering their results, their discussions around their findings and whether their discussions accurately reflected the reality of their data, with a keen focus on the presence of effect size measures.

2.1 Example of the Importance of Effect Size

Before beginning of discussions on effect size, an example may highlight why effect sizes are important and provide some context to why they should be used. One example that highlights the importance of reporting on effect size is smoking data collected from Arizona high school kids. The data appears as a 3x2 contingency table with factors relating to both the students smoking status as well as their parent's smoking status. This data set is tabulated below:

	Parent Smoking Status		
Student Smoking Status	Both	One	Neither
Smokes	400	416	188
Does not Smoke	1380	2239	1356

To test whether the smoking status of students is affected by the smoking status of their parents, we would be interested to see if the groups are independently distributed or not. Here independence would imply that the parents would have no statistically significant effect on their children, and non-independence would imply that there exists some relationship between those two factors.

To perform an independence test contingency data, we use a Chi-square test, using R we receive the following result:

χ^2 Value	Degrees of Freedom	Significance
37.566	2	6.959e-09

Here this suggests that the distribution of categories is not independent and thus there is some underlying relationship appearing in the data. However, when we visualise this data set, we see the following:

Figure 1: Smoking Data Mosaic

Now it is much easier to see that there are slight differences between each of the groups, however the overall relationship, or difference between each of the categories is not overly large. The chi-square test we used earlier is extremely sensitive to sample size and hence returned a significance level that was extremely small.

Without visualisation, it could be easy to present that parents smoking status impacts on their children's smoking status, without mention of the magnitude of the relationship between groups. Like most significance tests, Chi-square does not provide much insight into how big this relationship is, this is where the effect size measures come in. They help to measure the relationships that were found to be significant.
For categorical data, there exists many effect sizes, and as earlier it was mentioned each have their appropriate contexts. For this data, a 3x2 contingency table, Cramer's V proved the most appropriate measure.

Cramer's V is calculated as followed:

$$V = \sqrt{\frac{\chi^2/N}{min(k-1, r-1)}}$$

Where k and r are the number of factors. Calculating it for this data, we find Cramer's V is $V = 0.08360076$, this suggests that a very small relationship exists between the categories. This is again seen through the mosaic plot above.

3 Measures of Effect

Most effect sizes can be categorised into the following families:

- Correlation family
- Difference family
- Categorical family

Each group of effect sizes has its particular uses and each member of its family may be more appropriate for various data and/or experiment types.

3.1 Correlation family

Correlation family effect sizes aim to explore how much variation observed in the results can be explained by the independent variables. It can also be used to explore how closely related two variables are and how movement of one variable can affect movement in the other.

Most commonly known and used would refer to the coefficient of determination, denoted as R^2, which is commonly used as a 'measure of fit' for linear models. Reporting on how much of the variation in the outcome can be predicted by the linear model fitted. Other commonly used and reported on correlation effect measures include Pearson's r and η^2.

A brief example of correlation effect is amphipods data, in which counts of amphipods from locations was taken as well as other measures from their environment. We are interested in what is correlated with amphipod population sizes so we can understand what impacts their population. This data set includes a range of values for live seagrass, dead seagrass, drift algae and epiphytes.

As we are interested in understanding what factors have the largest impact on amphipod populations, we can use Pearson's correlation which returns a value between -1 and 1 to suggest not only how correlated two variables are, but also the direction of that correlation.

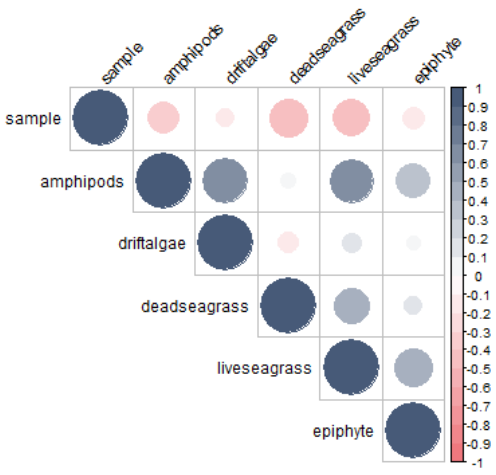

Figure 2: Amphipod Correlation Table

From here, it can be seen that live seagrass has a reasonably large correlation with amphipods, suggesting that higher amounts of live seagrass have a positive relationship with amphipod population sizes.

If we want to plot this relationship, we can see the following:

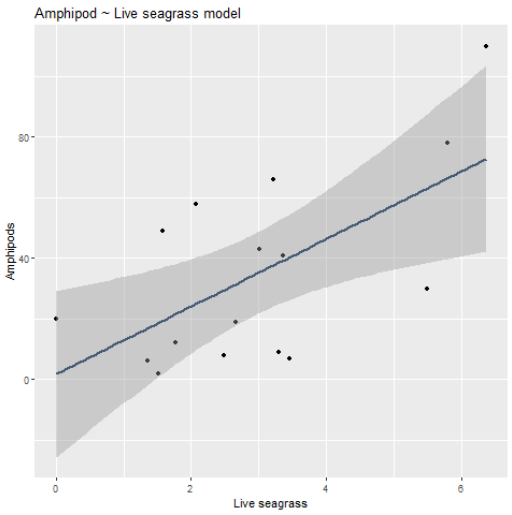

Figure 3: Linear model of Amphipods and Live Seagrass

Again, showing that there is a positive relationship between these variables. If we want to find the real value of correlation, we can calculate Pearson's correlation co-efficient directly, which is done as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}$$

Pearson's correlation coefficient can also be converted into a t-distribution test statistic, to find if the correlation is statistically significant or not, this is done by:

$$t = \frac{r^2}{\sqrt{1 - r^2}}\sqrt{n - 2}$$

Now collating both of these results we find the following:

t Value	Significance	Correlation Estimate
2.9581	0.01038	0.6201772

Which reading from here, it shows that there is a statistically significant correlation found, but like most statistical tests, it does not refer to how large or the direction of this correlation. However, the Pearson's coefficient here shows a value of 0.6201772 suggesting that roughly 62% of the variation in amphipod population size can be explained by variation in live seagrass.

In this example, we use Pearson's r to find variables with the most correlation in order to understand what environmental factors have larger impacts on amphipod populations. Understanding correlation between variables and performing correlation tests before creating linear models can help make more robust linear models.

3.2 Difference family

Difference effect measures typically aim to consider the differences between two or more groups' means. This is similar to performing a t-test on one's data, where a t-test will determine how statistically different the groups are based on their means, the effect size measure can tell how small or large this difference is.

As the t-test is sensitive to sample size, it can be important to report on the effect size difference between the groups in order to truly understand how different each group is. This is also an important part of analysis, as large samples can return statistically significant results, however its effect size should remain consistent and determine whether the size of the differences between the groups. Reporting on the effect size can provide more context in understanding what the data is truly representing.

Including another brief example of difference effect size measures, we have data from an experiment that took place to view the effect of magnetic fields on the flow of calcium out of chicken's brains. Two groups were observed, the first being a control group of 32 chickens, and the second being an exposed group of 36 chickens. The aim of this experiment was to observe the differences between these two groups.

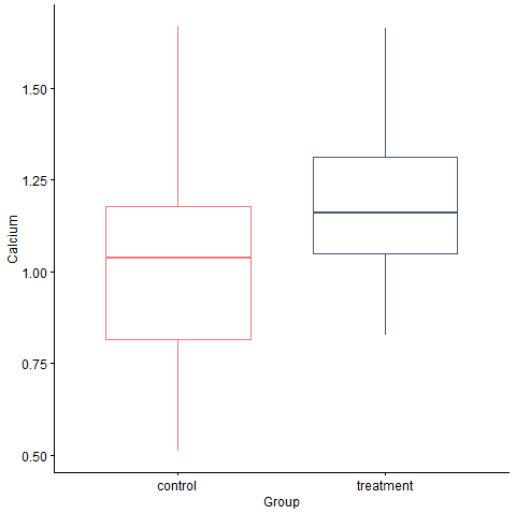

Figure 4: Box plot of Chicken Data

From here we can see that the groups are slightly different, however, to see if this is statistically significant, we can use the t-test to test the means of each of the groups.

t Value	DF	Significance
-2.9663	60.624	0.004308

Now, looking at the result from the t-test, we can see that the difference is statistically different, however we cannot determine if this difference is small or large. This is where the effect size measure comes in, here we will use Cohen's d, which is a commonly used effect measure. Cohen's d is calculated as follows:

$$Cohen's\ d = \frac{M_2 - M_1}{SD_{pooled}}$$

Where $SD_{pooled} = \sqrt{\frac{SD_1^2 + SD_2^2}{2}}$

From this data set, Cohen's d finds the value 0.6668541, which we interpret as being a 'medium' difference between groups. This can also be interpreted as the average score from the treatment group is 0.6668541 times the standard deviations above the average score from the control group.

For this data, the control group had a mean of 1.013 and a standard deviation of 0.24, if we want to test if the effect size does show how many standard deviations above the control mean the test mean is we can calculate:

$$1.013 + 0.24 \times 0.6668541 = 1.173045$$

Which is close the control group's real mean of 0.172.

In the difference effect size measure family, there are three measures that are used frequently, and are calculated in similar ways however they have their own usages.

Firstly, Cohen's d, which was used above, is a commonly used measure and is used for situations where the standard deviations of the two groups are similar and when the sample sizes of each group are similar as well.

The chicken data satisfies both of these conditions.

Next is Glass's Delta, which is used when the standard deviation of the groups vary drastically, or if you are comparing multiple treatment groups to the control group. It is calculated as follows:

$$Glass's\ \Delta = \frac{M_{control} - M_{experiment}}{SD_{control}}$$

Finally, Hedge's g, which is commonly used when sample sizes are smaller or if sample sizes vary. This is calculated very similarly to Cohen's d; however, the pooled standard deviation has a different weighting and therefore produces slightly different results. Hedge's g is calculated as follows:

$$Hedge's\ g = \frac{M_2 - M_1}{SD^*_{pooled}}$$

Where $SD^*_{pooled} = \sqrt{\frac{(n_1-1)SD_1^2+(n_2-1)SD_2^2}{n_1+n_2-2}}$

If we were to compare these three measures on the data set above, we find the following:

Cohen's d	Hedge's g	Glass's Δ
0.7284813	0.7201716	0.6668541

So, the measures return similar results, and each return that the difference found is 'medium'. However, Cohen's d remains the most appropriate for this situation.

3.3 Categorical family

Categorical effect sizes are used to measure relationships amongst categorical variables. This was seen earlier through the smoking example, however, here another measure will be used to show the different situations of the measures.

As discussed earlier, a common test for independence of categorical variables includes the chi-squared test, which compares values of each category with their expected value if independence is true. However, we have seen that the chi-square test is extremely sensitive to its sample size. As the sample size increases, chi-square test is much more likely to find smaller differences to be statistically significant, the chi-square also only comments on whether the relationship is significant rather than the size of the relationship.

Categorical data can follow many different types of construction, the first being as we saw earlier in a contingency table in which the frequency is tabulated across two (or more) factors. Some effect size measures useful in these situations include Cramer's V, which is useful to understand the distribution of categories and was used earlier in the smoking example. Another measure is Pearson's Contingency Co-Efficient, similar to Cramer's V focuses on distribution of categories, however, is more appropriate for larger experimental designs (such as 5x5 contingency tables). Finally, ϕ which is useful for when comparing binary classification categories, (i.e. male/female).

These measures all provide extra information on how large, and potentially important, the difference between the distribution of these factors may be.

4 Literature Review

4.1 Sampling Process

The literature review was developed in order to be able to sample areas of research where statistical analysis may be present. The areas of study chosen for this review included, psychology, commerce, education, biology and ecology. Next, a list of 10 highly respected journals from each field was organised, in which University of Newcastle had full access to.

When sampling, we were more interested in more recent journal papers and their approach to statistics. Therefore, we limited our samples to only include papers published in the last 2 years, from Jan 2018 onwards.

To choose the articles sampled, we used a random sample function from R which the number of journals was included and roughly 40% would be selected. This would also include some journals that were not appropriate for sampling usage, such as reviews, memorials and philosophical discussions. These were excluded from the count when finding how many papers to sample.

4.2 Provisional Results

4.2.1 Economics

From economics and commerce papers, a large proportioned of those sampled included some form of numerical/statistical data. While some papers were more involved in studying long term trends and discussing possible causes over time for these trends, others were more invested in model fitting.

In many studies, the only measures reported on came from the coefficients from linear regression outputs, occasionally there were also inclusions of two tailed means test.

Many of the journals were interested in finding relationships between different variables and understanding their links. However, a large proportion of economics papers focused on fitting a linear model in order to assess how each variable is connected. While with linear model fitting, they report on the R^2 measure, it was rarely done in a way with consideration of what the value truly represents.

There were many cases in which models were fitted with a large number of variables which lead to the model having an incredibly high R^2 value. It is known that more variables added will eventually lead to a model that has explained a large proportion of its variety, hence a large R^2 value. This is considered 'over-fitting' in which a model is so adjusted to its input data, it can be difficult for that model to provide valuable insights on new or unseen data. An easy way to overcome this could be to use the adjusted R^2 measure, which considers the number of predictors and adjusts to whether over-fitting may be present.

While it is possible to gain some understandings of the relationships and correlations between variables with linear model fittings, there are other approaches as well that could provide insight in a more useful way. One could be using one of the many correlation family effect sizes, particularly Pearson's p, which aims to measure the correlation and strength of the correlation of variables.

For the few studies that did include effect size, most common measures reported were the Pearson's r, Cramer's V and Spearman's p. Overall from the samples read and reported on, it appears that many economics papers under-utilise statistics and overlook the benefits that could come from using effect sizes.

4.2.2 Biology and Ecology

Many ecology and biology papers looked at involve discussions on how different environmental factors interact. While some papers include statistical or numerical data, the majority did not. And like economic papers, some ecology papers were concerned with trends over time, especially with relation to climate change.

As most papers with numerical data were interested in the interaction and relationships between their observations, ANOVA was a commonly used metric. The ones concerned with long term trends were concerned with model fitting and finding relationships and predictions that way.

From the models concerned with linear models, R^2 remained a commonly reported on statistic, however many papers lead discussions on the co-efficient and what the linear model co-efficient imply in relation to the model. Very rarely was the Pearson's correlation measure mentioned, which does have uses in linear models as it helps to measure how much variation caused in one variable could possibly be explained by the variation in another.

Many ecology papers also included interaction plots to show the differences between factors. While this can be a useful way to visualise the underlying relationships, it is not a statistically rigorous way to measure the interaction and effect that the variables have on one another. When comparing the way that factors impact on one another, useful effect size measures come from the categorical family. As the categorical effect measures are concerned with the difference in behaviour across different groups, it could provide valuable insights into the distribution of their population. Common measures include, Cramer's V, Pearson's Contingency Coefficient and ϕ.

Common measures reported in ecology papers include reporting on the output of ANOVA tables, such as the F statistic, which is used to determine a statistically significant difference. However, there could be an under utilisation of certain measures that could be useful, such as eta-squared (η^2) which allow a better understanding of how much variance each factor is responsible.

4.2.3 Psychology

Many psychology papers have a significant proportion of their paper dedicated to data analysis and exploration. Psychology papers sampled from also included a reasonable number of meta-analysis papers, in which statistical analysis is performed using results from many previously published papers to hopefully better understand their proposed research objective.

Due to the nature of meta-analysis being focused on statistical analysis, many included rigorous and thorough analysis and explanation on their findings. Many of the meta-analysis papers included effect size measures such as Cohen's d, Hedge's g, Odds Ratio and Cochran's Q. The implications of each effect size were also discussed in terms of what it truly meant for the context in which the data came from.

Many statistical papers appeared to be concerned with the differences between groups, whether it be control and treatment groups, affected populations and non-affected populations etc. Therefore, a majority included some form of difference of two means test. However, there were a few cases in which linear models were constructed and analysed.

Many of the other individual study psychology papers did also report on their found effect sizes alongside their data analysis. Most used measured included Cohen's d, Hedge's g which are common ways to explore the magnitude of difference between groups. Other measures that were discussed includes R^2, as well as a standardised beta metric.

The standardised beta metric is calculated by recalculating each of the variables by subtracting its mean and dividing by its standard deviation. This reformulates variables to have a mean of 0 and a standard deviation of

1. Then a linear model is constructed in which the co-efficient now represent the 'relative importance' of each variable. The numerical value represents how many standard deviations the dependent variable will change, for an increase in the standard deviation of the independent variable.

Due to the standardised nature, this measure can be skewed when dealing with non-normal data. Again, it is also measuring how much difference in one variable's standard deviation will create a difference in another variables' standard deviation, it can be potentially misleading to report on as the scale of the original data can be lost. Another note is that when constructing a linear model, the variables can have some level of overlap, which may not be accurately portrayed through the co-efficient.

Essentially, standardised beta can provide some insights to linear model fitting, however, due to the fact it is influenced by the overlap that some variables may have on the independent variable, it cannot be considered conclusive in measuring effect, especially when other measures exist. Measures that could be useful for constructing linear models include Pearson's r, finding how closely linked variables may be, other measures that can be useful in comparing the variation explained by each variable include η^2 and ω^2 measures, as well as Cohen's f^2, which calculates the proportion of variance explained by each of the dependant variables.

Many of the papers sampled from psychology did include statistical analysis to support their research and back their findings. However, while a considerable proportion included sophisticated levels of statistical analysis and provided discussions on the implications of their findings, there was still a proportion that overlooked effect size and the role that it has in conceptualising their findings.

4.2.4 Education

In education studies, papers were concerned in factors that play into people's performances in education settings. A handful of the education papers sampled were also meta-analysis papers and thus had a more rigorous analysis of statistics. Many education papers also included philosophical discussions surrounding education and thus had no data analysis.

Some meta-analysis papers in this area chose to summarise the findings across many papers rather than analyse the data that was found from the individual studies. This could again be, possibly, that education measures across countries and cultures may not be entirely comparable or equal, thus making comparisons between results from different papers unsuitable.

A different effect size was discussed in education papers, namely Tau-U, which allows for the calculation of effect size where there exists only a single case of data to analyse. This could be a valuable metric to use as education studies may be difficult to perform, as they could have potential to disrupt their studies and may not be appropriate to perform often.

Tau-U is used for single case analyses, as in recording the performance before a treatment is started, then recording performance after a treatment is administered. This could be reflective of education studies as they appear focused on how individuals react, rather than differences between groups. Tau-U aims to remove the baseline trends in order to understand the trends that are apparent in both the 'control' period as well as the 'treatment' period in order to see if treatment has any effect on performance.

There are some substantial downfalls to this method of analysis, due to its construction, it is not bounded between -1 and 1, and thus it can make attempting to interpret the size of the effect (i.e. into groups such as 'small', 'medium' and 'large') much more difficult as well as an inability to be able to discuss these measures to the same guidelines that other measures have. Another limitation includes the lack of consistency with formulas and notations when discussing and calculating Tau-U measures , which can make comparisons across papers more difficult.

The more common effect size measures that were reported on from education papers include Cohen's d, Pearson's correlation and Tau-U. While they were reported on, it did not appear to be discussed in any substantial way relating to the research that was being performed. Many of the education papers that were sampled were not appropriate, and hence a smaller sample of these papers were included when reviewing effect size measures across fields of studies. However, the papers that did include analysis did leave room for deeper discussions on their findings and how it directly relates to the data at hand.

4.3 Fail-Safe N

Many psychology papers sampled were meta-analysis of other published psychology papers. Whilst many of these meta-analysis papers were more thorough when it came to effectively calculating effect sizes and discussing their interpretation and meaning, measures frequently referred to included Cohen's d, Hedge's g and Cochran's Q.

While meta-analysis provides insights from a range of studies and calculates appropriate effect measures, it can overlook a key factor when viewing only published papers. It is much more likely for papers that find statistically significant results to be published, meaning that during meta-analysis, whist looking through multiple papers, could also be missing a proportion of results that did not return statistically significant results.

A proposed method to overcome this issue, is to calculate the "Fail-Safe N", which represents the number of papers needed to reduce the meta-analysis results to non-significance. This can be useful to know, as if the returned "Fail-Safe N" is small, only a small number of non-significant results are required to reduce the effect size to non-significance. This could imply that possibly the result is not as good as it is framed to be, or that publishing bias is playing a role. However, a large "Fail-Safe N" is indicative of a result and effect size that would be harder to reduce to a non-significant result, and thus is more likely to be an accurate portrayal of the a relationship existing.

Like effect size measures, there exists multiple ways to calculate the Fail-Safe N, one way proposed by Rosenthan is as follows:

$$N_{fs} = ({}^{N_0}/_{Z_c^2})(N_0 \bar{Z_0}^2 - Z_c^2)$$

Where Z_c is the critical value of Z, N_0 is the number of studies looked at, $\bar{Z_0}$ is the average Z value found across all studies and N_{fs} is the number of non-significant papers are required to reduce the findings to non-significance.

One downside of this measure is that is although it is consistent with intuition, it does not have a basis in a statistical model. However there have been further developments in models and many proposed ways to calculate such statistics.

5 Power Analysis

When preparing to perform statistical analysis, one should consider their desired power, effect size that they would be interested in finding, their accepted statistical significance level as well as their sample size. These four components make up the basis of many statistical tests and interact in frequently conflicting ways. Therefore, understanding their relationships can improve the way that one approaches their statistical analysis.

As described earlier, statistical significance level, refers to the level that one rejects that the relationships apparent in the data are due to just chance. Most typically, many use a significance level of 5%, meaning there is a calculated 5% likelihood that the results are due to just chance and not any underlying phenomena.

Power is defined as the likelihood of the statistical test will reject the null hypothesis when the alternative hypothesis is true. I.e. finding the likelihood of saying that a relationship exists when one truly does exist.

The effect size is again, the measure of the relationship apparent in the data. Sample size refers to the number of data points that you have to perform the analysis with.

As significance finds the likelihood that the data is due to chance and hence is used to determine whether to accept or reject the null hypothesis and power is concerned with correctly rejecting the null hypothesis, these two have a direct relationship. If one wanted to increase the power of a statistical test, they could increase the level of significance. Making the statistical significance level higher means that there is more likelihood of rejecting the null hypothesis and thus will have a higher chance of rejecting the null hypothesis when it is false (and also rejecting the null hypothesis when it is true).

As shown before, many statistical tests are shown to be more likely to return a statistically significant result when the sample size is large. This is due to the effect size being easier to detect and thus show that the null hypothesis is not true. This can be used to imply that an increase in the sample size leads to an increase in the power of the statistical test all things being equal. An extension, as effect size grows, a smaller sample size will become more able to find a statistically significant result, as the relationship will still remain apparent even with smaller samples.

As the effect size one is interested in becomes smaller, smaller effect sizes can still be important in particular areas (especially medicine), more samples will be required in order for the statistical test to detect that the smaller effect size is a statistically significant relationship within the data. Similarly, if one already has a large sample size, the statistical test will be more likely to return a significant result, regardless of if one is interested in smaller effect size. So even if a large number of samples are collected, the result may not actually be substantial when interpreting its real world meaning through effect size. This should be a consideration when performing experiments and collecting data to ensure that the data collection is appropriately designed in order that its desired effect size and significance level can be achieved through its desired power and sample size.

6 Simulation

As mentioned early, one reason the effect size is important is that it is able to overcome some of the limitations that many significance tests have. The main limitation being that if given a large enough sample size, very small differences can become statistically significant.

Many effect sizes are believed to be more robust when it comes to studies of larger sample sizes. In order to test this claim, a simulation study will be prepared in which many different effect sizes will be tested against generated data from many distributions. We will be exploring both the statistical significance and effect size measures and their behaviours to better understand them.

6.1 Correlation Family

For testing correlation measures, two sets of data were created to have a specified correlation between them. In R, this can be done easily using the 'mvrnorm' function. As we are interested in the long term behaviour of the measure, the sample sizes that were used started at 25 and double each step until 51200, making 12 different sample sizes, then having 1000 iterations at each sample size, in which new data was generated, the correlation effect calculated (Pearson's r, R^2 and η^2) and each stored in a row in a data frame.

This report will briefly explore 'small', and 'medium' correlations and see if the results act in a similar manner for each simulation.

For a 'small' correlation effect, the accepted baseline for correlation is 0.1. When generating values, x was generated with a mean of 1 and a standard deviation of 1 whilst y was generated with mean of 3 and a standard deviation of 1 as well. The mvrnorm function was used, and the given correlation value was 0.15.

As the formulation of these measures are different, we will be observing Pearson's r on a separate graph as it varies from -1 to 1 and measures the slop of the linear relationship that exists between the two variables. The other measures, R^2 and η^2 aim to estimate how much variance in the independent variable can be explained by variance in the dependant variable.

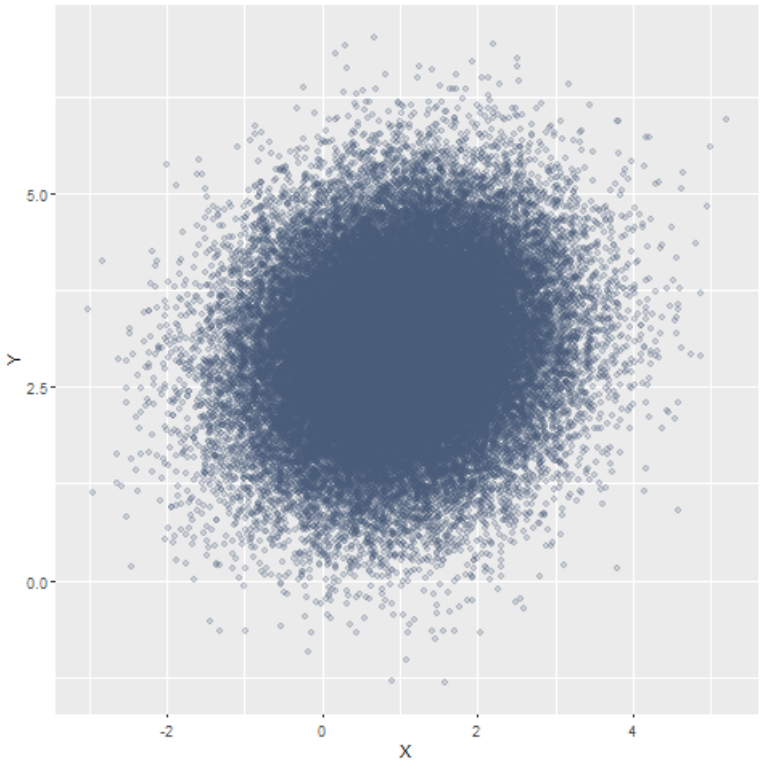

Figure 5: Sample of Simulated Data

Observing the estimates of both of these measures, we can see the following:

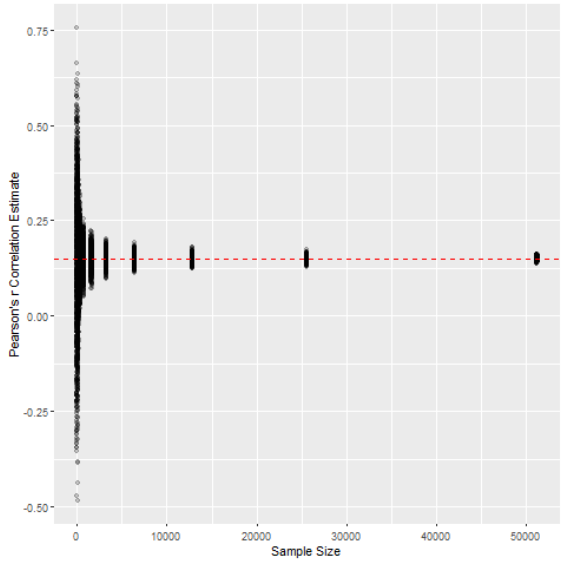

Figure 6: Pearson's Correlation Estimates

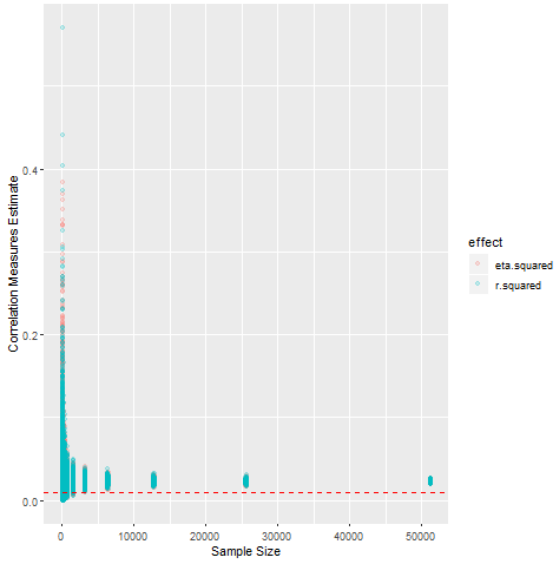

Figure 7: R^2 and η^2 Estimates

From here, it can be seen that each of the estimates produce a large variety of values at smaller samples sizes before converging onto a value at larger samples. The left hand side shows the Pearson's Correlation and the red line represents its generated correlation of 0.15, the right show the other estimates, R^2 and η^2 in which the dashed line their represents a 'small' benchmark value. I.e. only a small amount of variation in the in dependant variable is explained by the dependant variable.

Since we are interested in the behaviour of these measures as the sample sizes increase, we can observe both the mean of the estimates at each sample size as well as the standard deviation of calculated estimates.

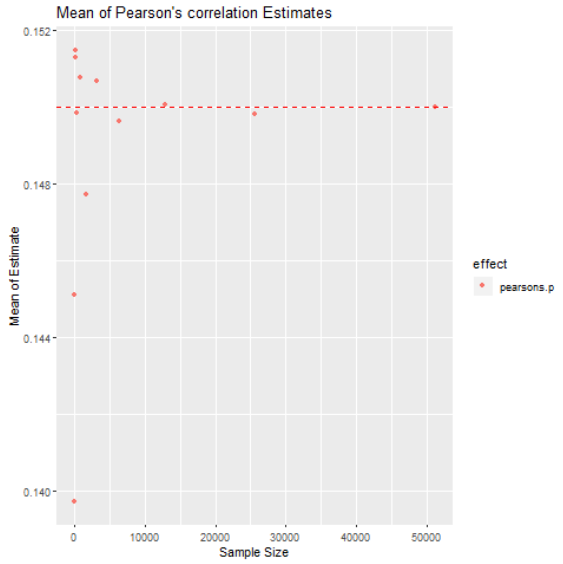

Figure 8: Mean of Pearson's Correlation Estimates

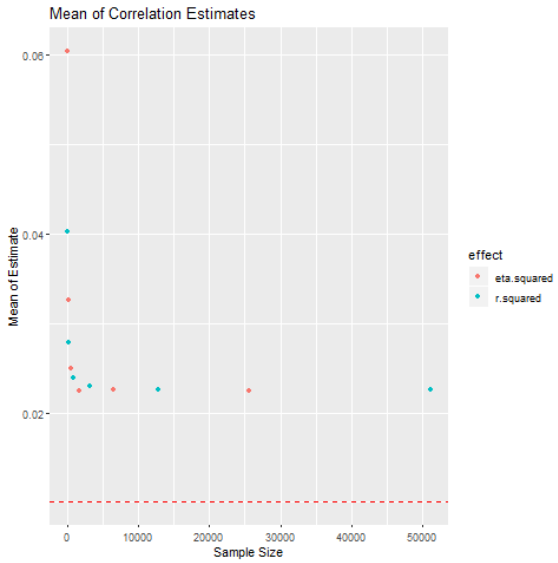

Figure 9: Mean of R^2 and η^2 Estimates

Next plotting the standard deviations of the measures:

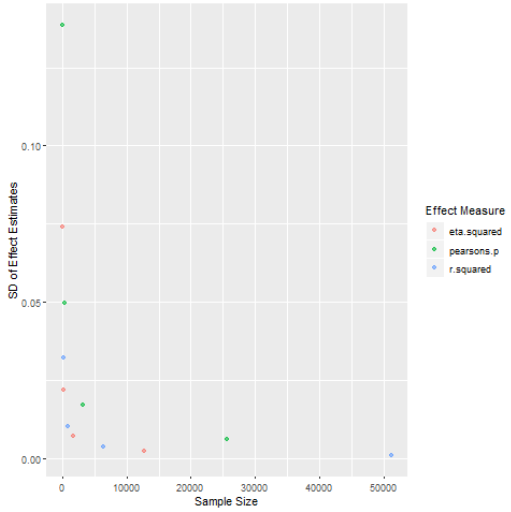

Figure 10: Standard Deviation of Measure Estimates

From here, the relationship we saw earlier is easily seen, as the sample size grows, the standard deviation of the estimates of each measure can be seen to decrease increasingly faster at each sample size.

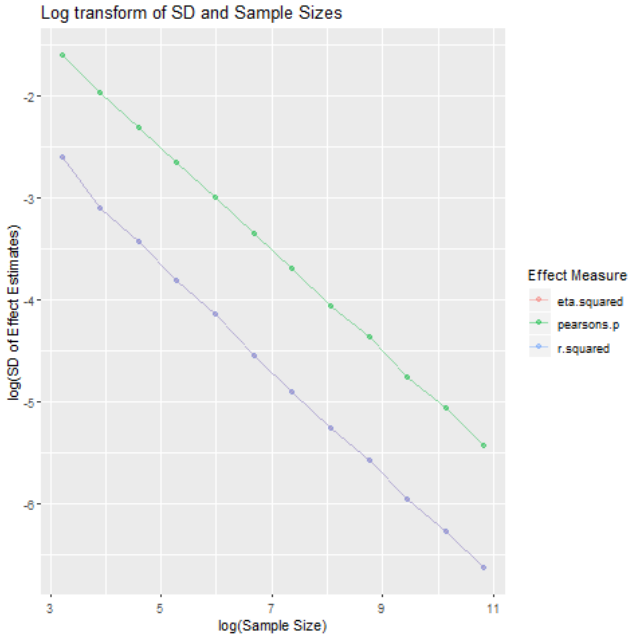

Figure 11: Log Transformation of Standard Deviations and Sample Size

From here, we can see that a linear relationship exists between the log transforms. This suggests that the relationship between the sample size and standard deviation. If we fit a linear line to this relation we find the slope of this line to be -0.5. This suggests that the relationship that exists between sample size and standard deviation of measures is inverse square. Meaning that for each increase in the sample size, the variation across the variables get increasingly smaller. This would also suggest that it approaches the population value with increasing accuracy as the sample size grows.

Now to see if this result remains consistent across different means, variances and effect sizes. Next generating data with means of 4 and 8.5 and standard deviations of 1.2 and 3.2 respectively. Using mvrnorm, we will set the two variables to have correlation of 0.35.

As the formulation of these measures are different, we will be observing Pearson's r on a separate graph as it varies from -1 to 1 and measures the slop of the linear relationship that exists between the two variables. The other measures, R^2 and η^2 aim to estimate how much variance in the independent variable can be explained by variance in the dependant variable.

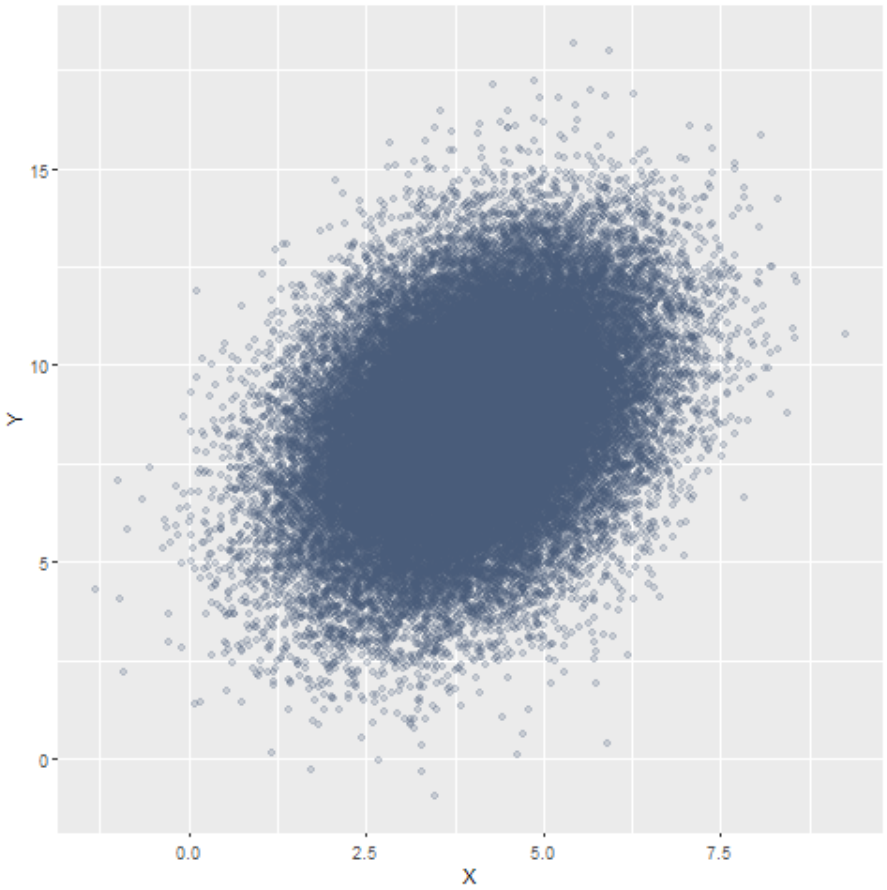

Figure 12: Sample of Simulated Data

Observing the estimates of both of these measures, we can see the following:

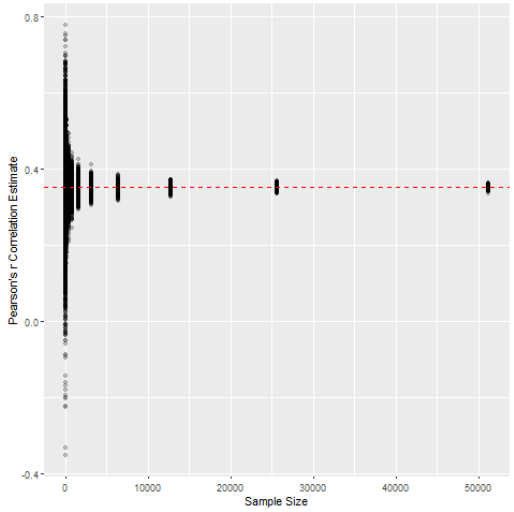

Figure 13: Pearson's Correlation Estimates

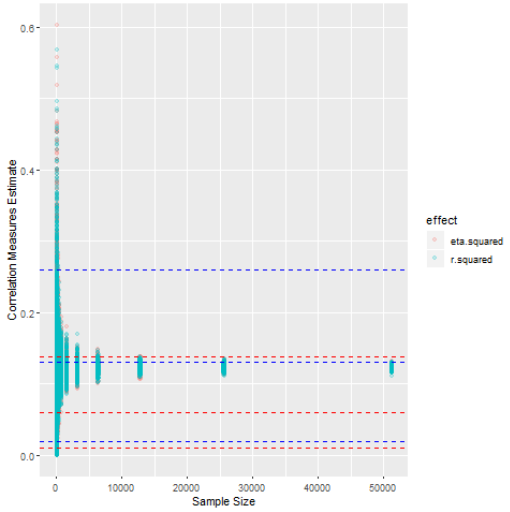

Figure 14: R^2 and η^2 Estimates

Again, it is seen that each of the estimates produce a large variety of values at smaller samples sizes before converging onto a value at larger samples. The left hand side shows the Pearson's Correlation and the red line represents its generated correlation of 0.35, the right show the other estimates, R^2 and η^2 in which the dashed line represent the boundaries between small medium and large. These measures work differently compared to Pearson's which aims to find how linearly correlated the two variables are, and how much an increase in one can lead to an increase in the other. R^2 and η^2 are more concerned in the amount of variability in the independent factor can be explained by variation in the dependant variable. For building a statistical model, it is important to be mindful of both of these measures in order to understanding how the variables are interacting with each other.

Again, as we are interested in long term behaviours, we observe both the mean of the estimates at each sample size as well as the standard deviation of calculated estimates.

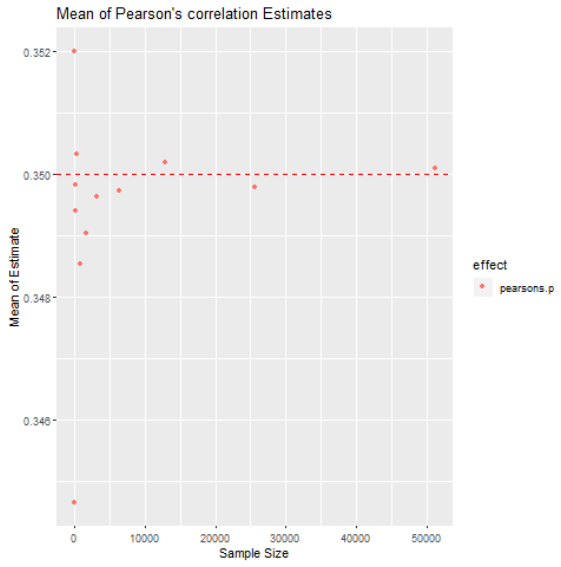

Figure 15: Mean of Pearson's Correlation Estimates

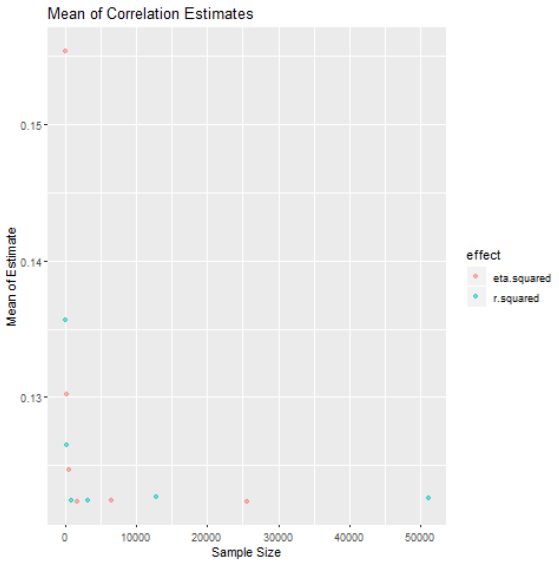

Figure 16: Mean of R^2 and η^2 Estimates

Next plotting the standard deviations of the measures:

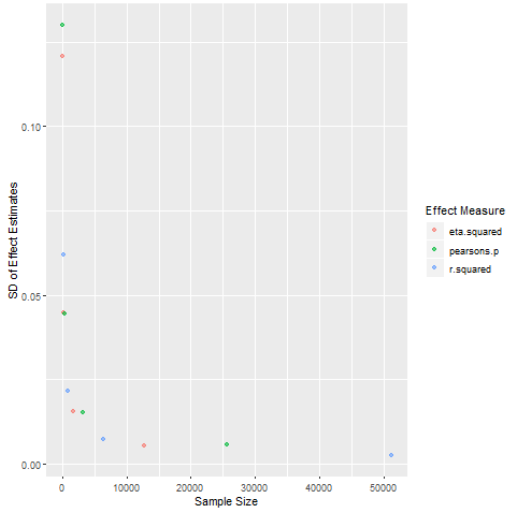

Figure 17: Standard Deviation of Measure Estimates

From here, the relationship we saw earlier can be observed again, as the sample size grows, the standard deviation of the estimates of each measure can be seen to decrease increasingly faster at each sample size.

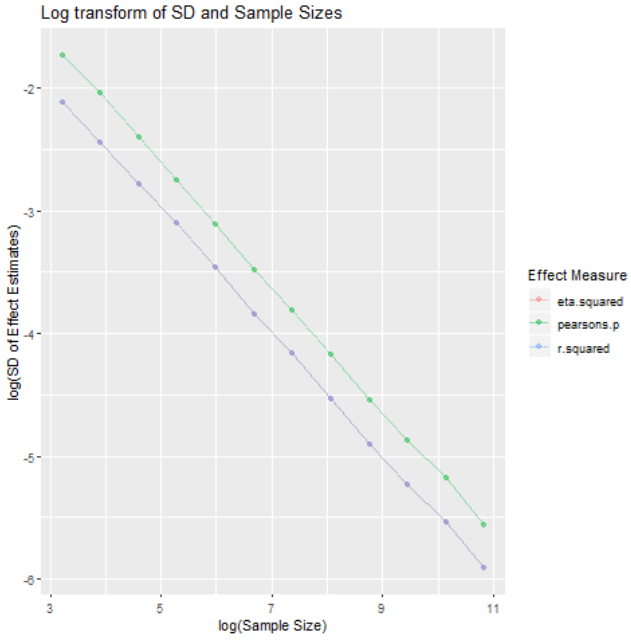

Figure 18: Log Transformation of Standard Deviations and Sample Size

Again it can be observed that the log transformations create a linear model in which the slope is calculated to be -0.5, again suggestive of an inverse square relationship.

6.2 Difference Family

Similar for difference measures, we are interested in the long-term behaviour of the measures and how they perform through increasing sample sizes. Much like the Correlation measure

simulation, sample sizes the data was generated at started at 25 and doubled each step ending at 51200. The data generated here also included 2 variables generated from normal distributions (using rnorm), however each group had a different given mean and standard deviation for the effect measure to estimate the difference between the two groups. For each sample size, 1000 simulations were performed in which samples for both groups were created, effect sizes calculated and stored before taking the averages across each sample size and measure. The effect measures that were used in this simulation include Cohen's d, Glass's Δ and Hedge's g.

For simulating a 'large' effect size difference between the groups, data was generated such that x has a mean of 7.16 and a standard deviation of 3 and y has mean of 4.6 and standard deviation of 3.2. This would calculate that the mean of x is roughly 0.8 standard deviations above the mean of y, and thus a large difference between the groups.

Plotting the difference estimates against the sample sizes, again shows that as the sample size grows, the effect measure estimates become closer together converge towards the real value. For this data, as the groups had two slightly different standard deviations the real effect size is slightly larger than 0.8 which informed the values we picked. The red line marked on this plot is at 0.8 which is what Cohen defines as the general guideline on a 'large' difference between groups. -

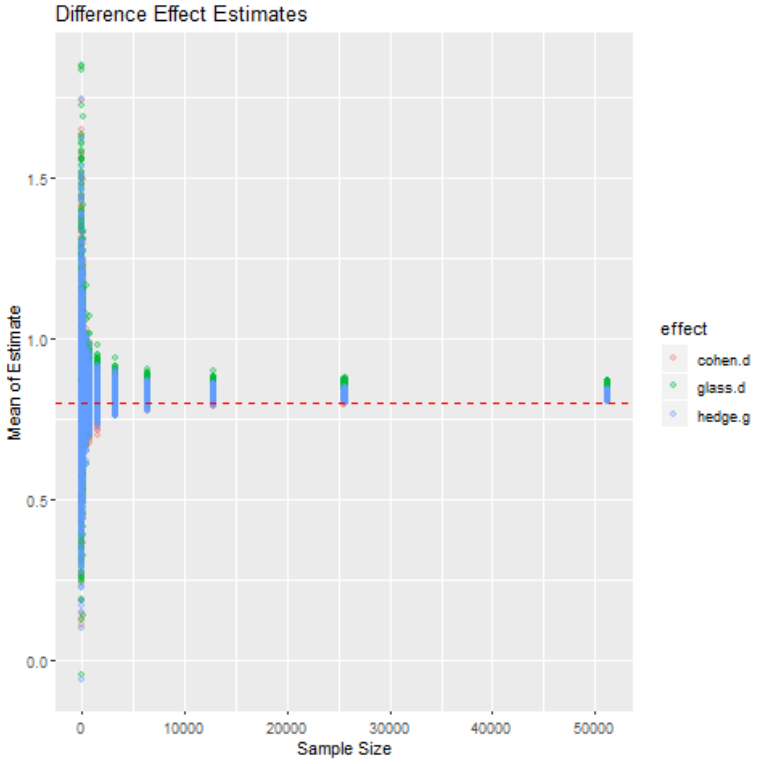

Figure 19: Effect Size Estimates

Following from the raw estimates, we can analyse the behaviour of the measures by taking the average of 1000 simulations and 12 sample sizes using 3 measures. Below, the left graph depicts the mean estimate of each measure and the right depicts the standard deviation across those 1000 values.

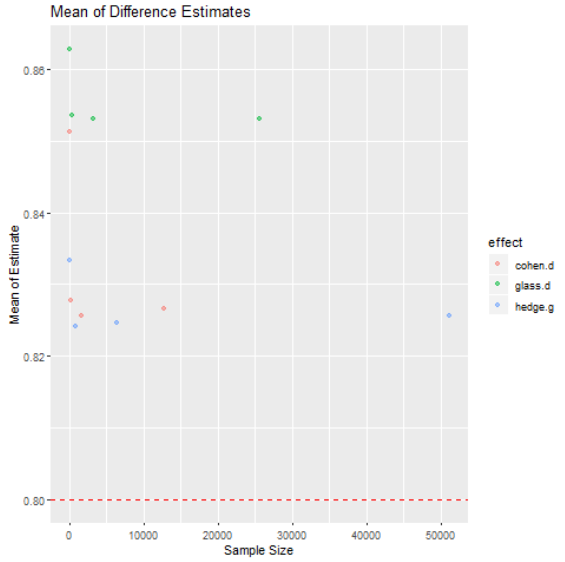

Figure 20: Mean of Estimates

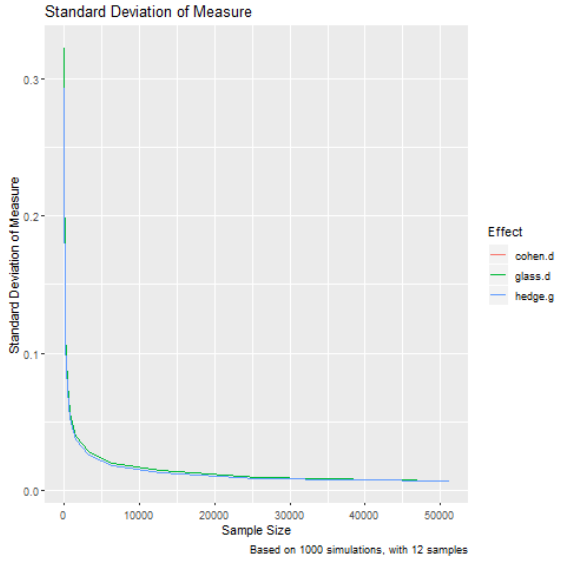

Figure 21: Standard Deviation of Estimates

We can see that all measures reduce the standard deviation in their estimates as the sample size increases. This suggests that as more data is available to these measures, they become more likely to calculate the populations real effect size. If we perform log transformations on both sample size and the standard deviations of the measure, we can observe a linear relationship between the two.

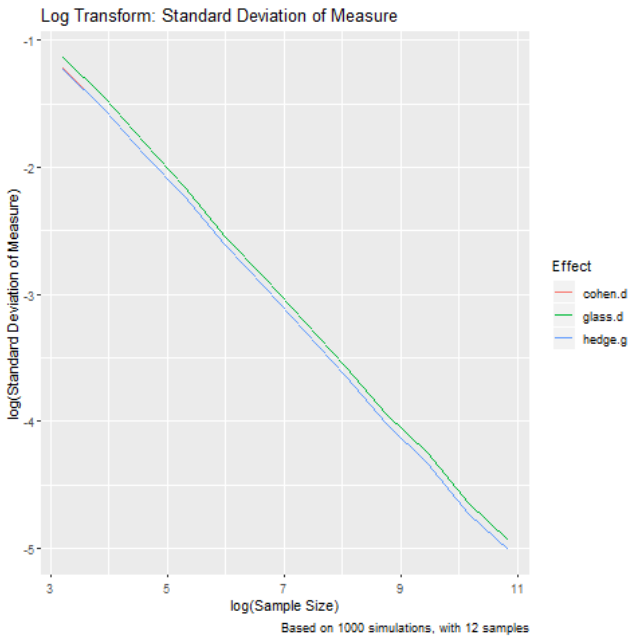

Figure 22: Log Transform of Sample Size and Standard Deviation

For simulating a 'small' effect size difference between the groups, data was generated such that x has a mean of 1.83 and a standard deviation of 5 and y has mean of 1.05 and standard deviation of 5.3. This would calculate that the mean of x is roughly 0.256 standard deviations above the mean of y, and thus a 'small' difference between the groups.

Again plotting the difference estimates against the sample sizes, shows that as the sample size grows, the effect measure estimates converge towards the real value.For this data, the guidelines proposed for Cohen suggests that measures of roughly 0.2 are considered ' small', and hence the line marked here represents that cut off.

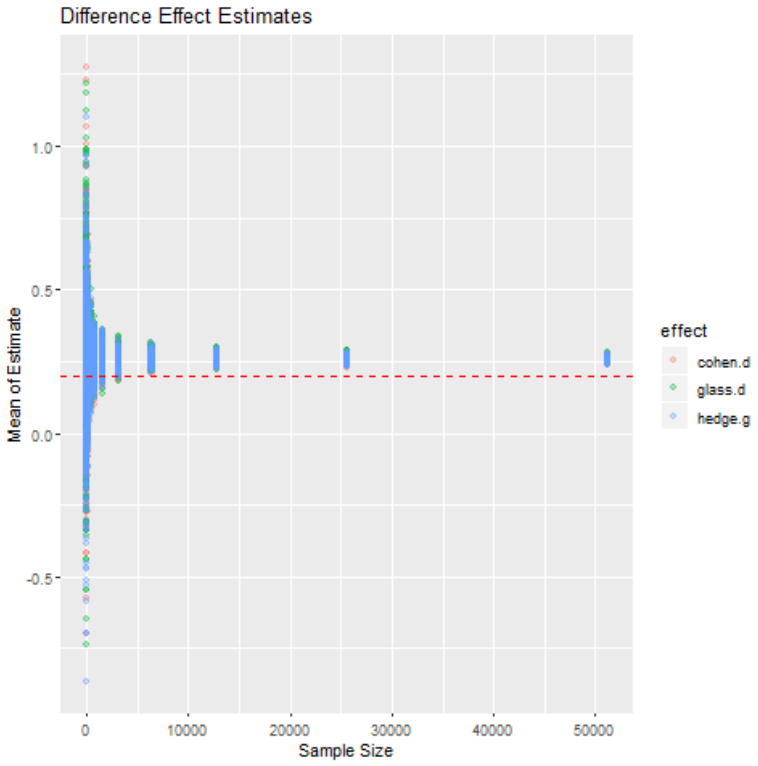

Figure 23: Effect Size Estimates

Similarly, observing the behaviour of the measures bu finding the means and standard deviations of the estimates we can find similar patterns. Below, the left graph depicts the mean estimate of each measure, where the red line represents Cohen's 'small' benchmark and the blue line represents the 'real' population effect size. The right depicts the standard deviation of the estimates across each sample size and measure.

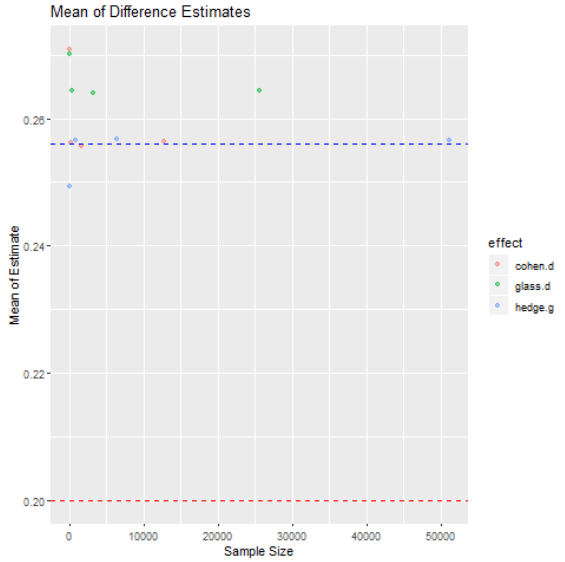

Figure 24: Mean

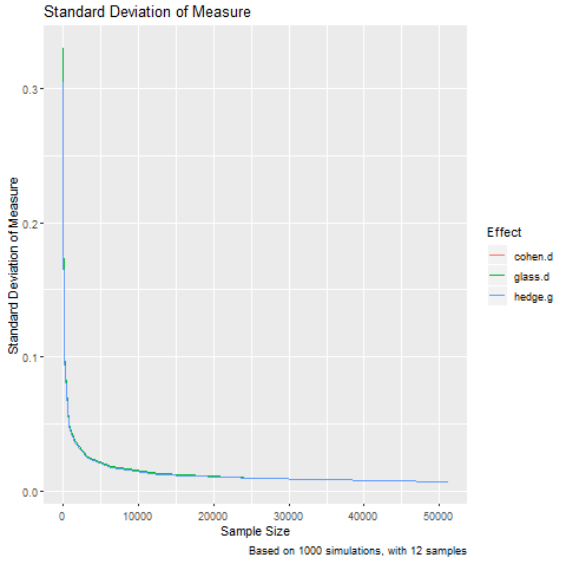

Figure 25: Standard Deviation

We again can see that all measures reduce the standard deviation in their estimates as the sample size increases. Again, this can be represented as a linear relationship through log transforming both the sample size as well as the standard deviation. Which calculating the slope of both this linear relationship as well as the found earlier, the slope of -0.5 again suggests that there is a power relationship between the two, and that the relationship present is an inverse square relationship. This again suggests that any increase in the sample size will lead to a larger reduction in the standard deviation of the measure and thus, an effect estimate that is more likely to be in line with the population's true effect.

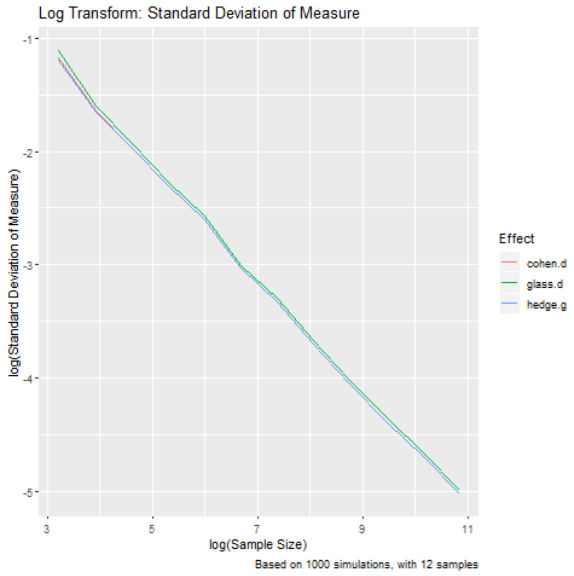

Figure 26: Log Transform

6.3 Categorical Family

Categorical measures were simulated in a different way. As we will be interested in the way that different groups interact, groups were simulated where a given probability for each group was given, next the simulated groups were arranged in a contingency table form, in this simulation we will be focusing on 3x2 experimental design.

The same sample size set up was used as previously in which starting with a small sample of 25, doubling at each step, however due to the limitations of the computer, ending with 12800. Then from the initially generated table, sampling was taken, with replacement allowed. This allows for weighted choices to be made and can also reflect what may happen when taking a real survey. One is more likely to record results from the already larger groups.

Firstly, showing a small difference in the distribution of groups, the image below shows the initial data in which the samples were taken.

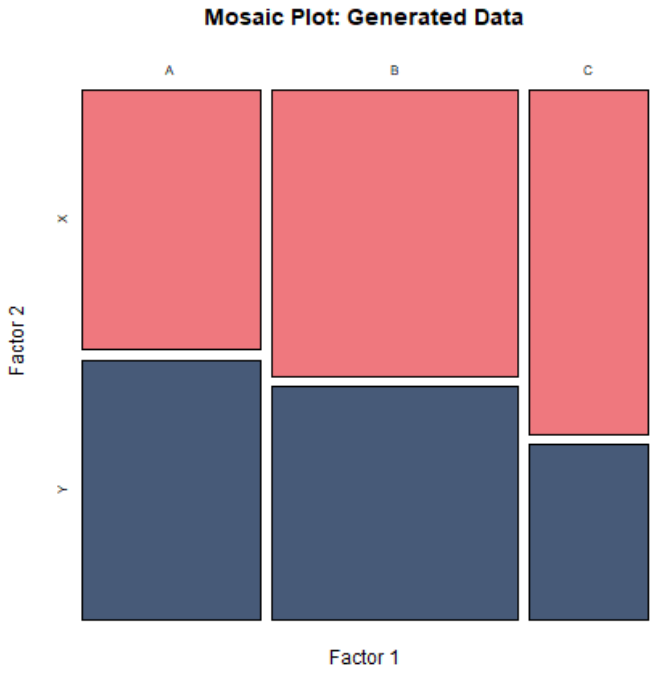

Figure 27: Original Simulated Distribution

After iterating through each sample size 150 and recording each of the estimates, plotting the estimates against their sample sizes we see a similar pattern as before, in which the measures start with a larger spread and more variation as the sample size is small and appear to converge towards a value as the sample size grows, here the dashed line represent the rough guidelines on 'small', 'medium' and 'large' effect sizes.

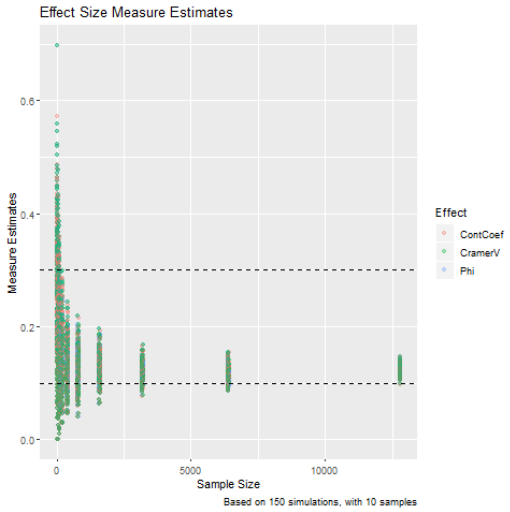

Figure 28: Categorical Effect Size Measure Estimates

If we want to view both the mean and standard deviations for each of the measures, we can see them below:

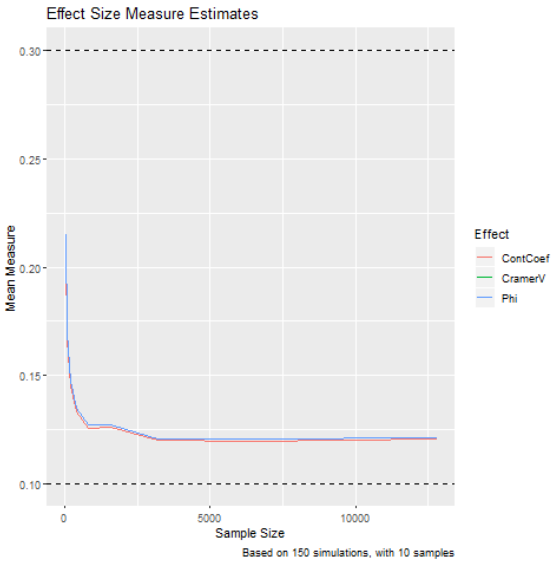

Figure 29: Mean

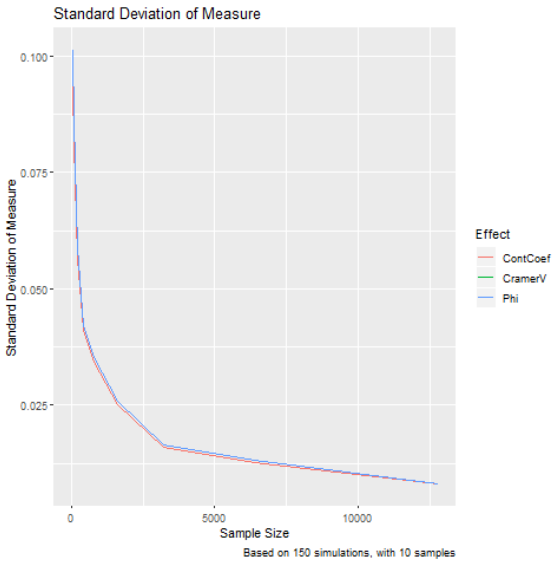

Figure 30: Standard Deviation

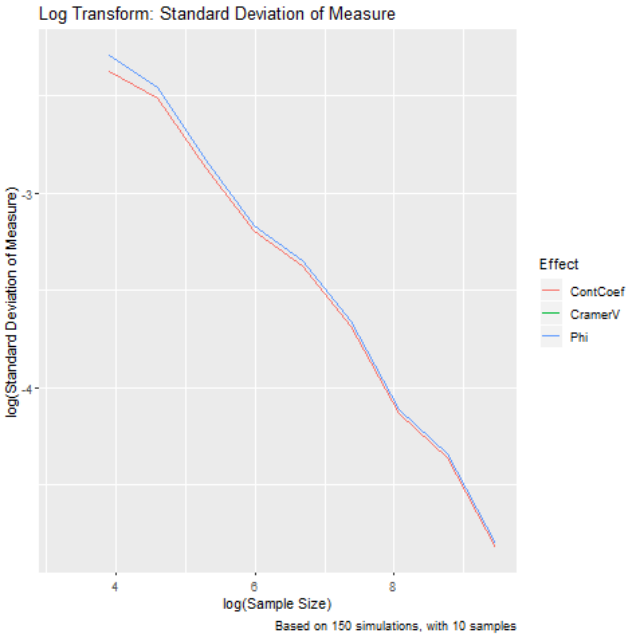

Figure 31: Log Transform

Again, we can see that a log transform can be applied on both the sample size and the standard deviation of the effect estimates to produce a linear relationship.

If we were to repeat this with categories that had larger differences between them we would find the following, starting with our new sampling distribution.

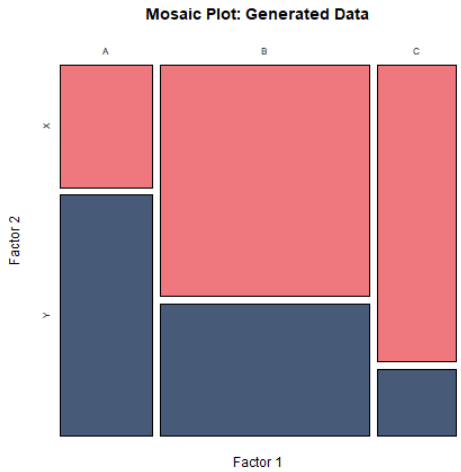

Figure 32: Original Simulated Distribution

Again, repeating 150 times at 10 different sample sizes, we will gain the following estimates, which as we can see now is above the 'medium' effect size line, suggesting that there is a relationship between the factors. As Cramer's V is concerned with the correlation between the factors, a larger affect size implies that there is a larger association between the factors

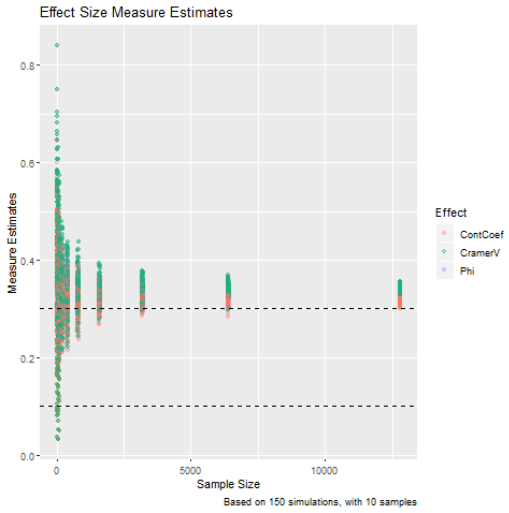

Figure 33: Categorical Effect Size Measure Estimates

Viewing both the mean of the estimates as well as the standard deviations of the estimates, we find similar patterns, however, comparing just to the last graphs, it appears that the mean of the estimate reaches the estimate faster, i.e. smaller sample sizes. This again backs up previous discussions from the *Power Analysis* section which acknowledges how larger effect sizes can be detected in smaller sample sizes.

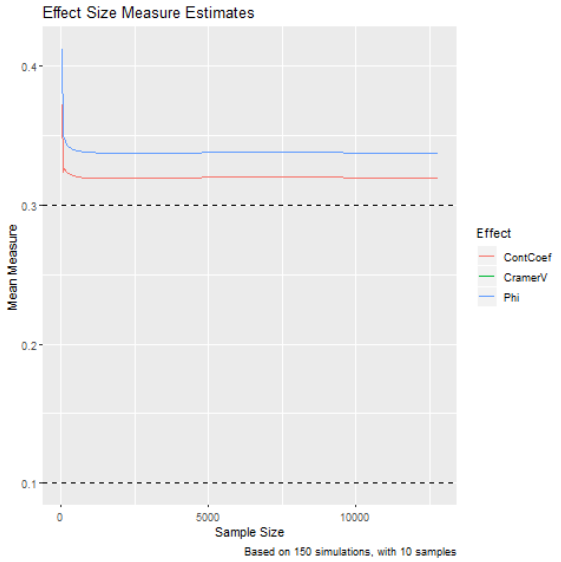

Figure 34: Mean

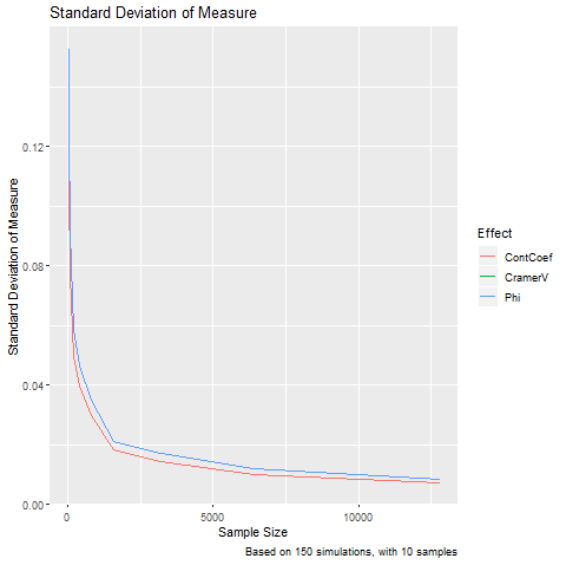

Figure 35: Standard Deviation

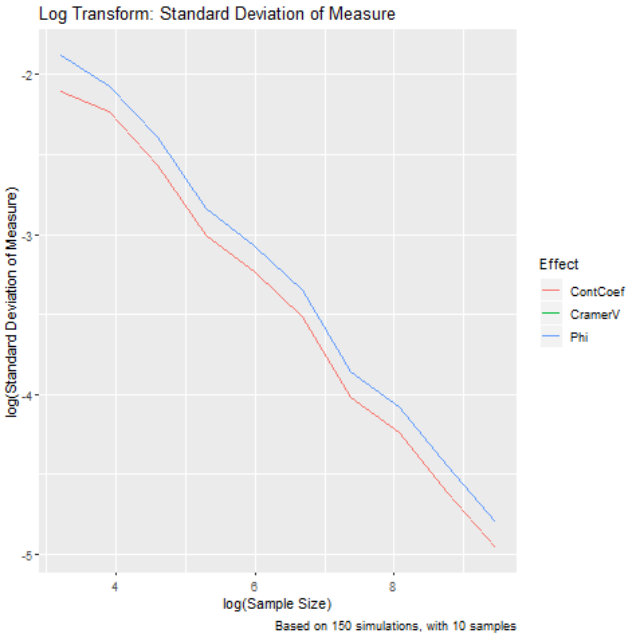

Figure 36: Log Transform

Performing log transforms on both sample size and standard deviation, we again find a linear relationship exists.

This furthermore suggests that regardless of the size of the effect, the standard deviation and variability of the measures will decrease increasingly as the sample size grows. Another important note, which can be seen in the estimates graphs on both pages, is that the Cramer's V measure is likely to over-estimate the value of the association between the categories, however this diminishes quickly as the sample size increases.

7 Conclusion

In hypothesis testing it can only be said if there is an effect, depending on the acceptance of null hypothesis, the effect size extends the limitations of significance testing and allows to measure the effect on populations. Effect size measures can provide many insights towards better understanding the relationships that exist within data sets. Effect measures are also able to measure both the magnitude and direction of particular relationships.

Effect size measures bleed into every aspect of statistical testing. This becomes apparent when considering how to prepare ones experiment in order to reach a satisfactory level of power, statistical significance while considering the magnitude of effect size that would be important in that particular scenario.

Through the literature study, it was seen that while many fields have components of statistical analysis, and many seem comfortable with data analysis techniques such as linear regression, ANOVA and finding the difference of two means, a lack of effect size reduces how much insight can be gained from their analysis. While meta-analysis does provide some rigour when dealing with statistical tests, many papers could be enhanced through a thorough discussion of the implications of their findings through their real world meanings. As such is possible when reporting on the effect size and its interpretation in relation to the data analysed.

There are many different types of measures, ranging from their different family groups, as well as multiple within each group that have various uses, appropriate contexts and limitations. Like many statistical tools, effect size measures do have some limitations, however, a reasonable understanding of the measures will allow for appropriate usage and thus a deeper understanding of one's data. This paper aimed to introduce the concept of effect sizes, show their usages, discuss some limitations, but overall bring attention to an often underutilised statistical concept.

8 References

Coe, R. (2002). It's the effect size, stupid: what effect size is and why it is important. Retrieved from https://www.leeds.ac.uk/educol/documents/00002182.html

Cohen, J. (1988). Statistical power analysis for the behavioural sciences. Hillsdale, NJ: Erlbaum.

Driscoll P, Wardrope J. An introduction to statistics. J Accid Emerg Med2000; 17:205.

Jones, S.R., Carley, S. Harrison, M. (2003). An introduction to power and sample size estimation, Emergency Medicine Journal, 20:453-458.

Orwin, R. (1983). A Fail-Safe N for Effect Size in Meta-Analysis. Journal of Educational Statistics, 8(2), 157-159. doi:10.2307/1164923

Vidgen, B. Yasseri, T. (2017). P-values: Misunderstood and misused, Retrieved from http://pantaneto.co.uk/p-values-misunderstood-and-misused-bertie-vidgen-and-taha-yasseri/