BSHDS2

Statistics II CA2

Christopher Weir

X19317131

[X19317131@student.ncirl.ie](mailto:X19317131@student.ncirl.ie)

One of the datasets I have chosen for this assignment is movie ratings from 2002-2016. The data has been sourced from Kaggle. (Allen, 2017) This dataset contains over 200 rows (movies) and 19 columns ranging from multiple genres to box office to movie ratings and age allowance. Each movie has an age rating, 4 critic reviews, their box office numbers, the year it was released and which genre(s) it is categorised as. Genres that are included are Horror, Children, Animation, Comedy, Drama, International among others and some movies only have 1 genre but some have multiple genres. As age rating system is American, which ranks from PG (Parental Guidance), PG13 (Parental Guidance for 13+) and R (Rated 18+).

**MANOVA**

Firstly, I am going to set out the null and alternative hypothesis.

**H0 = There are differences in the critics’ ratings of movies depending on the age rating of the movie**

**H1 = There are no differences in the critics’ ratings of movies depending on the age rating of the movie**

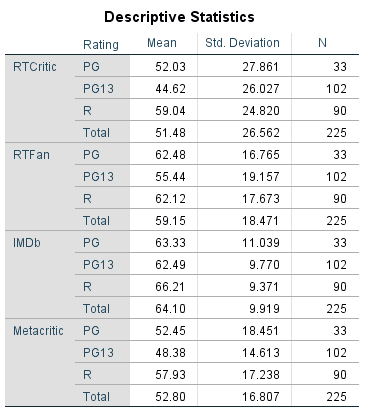


Figure : Descriptive Statistics of Variables tested in dataset

It can be noted that PG13 movies tend to have the worst average rating across all critics’ platforms with the standard deviations changing with each critic. However, each critic has in and around the same standard deviation for all 3 age ratings.

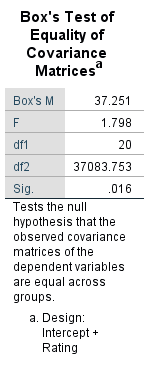


Figure : Box's Test of Equality of Covariance Matrices using the independent variable as the design intercept

From this test above, the box’s M value of 37.251 has a p-value of 0.016. As this is less than 0.05, we can conclude that the differences between the means are statistically significant which makes it eligible for a one-way MANOVA.

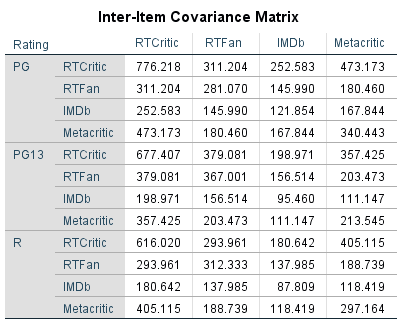


Figure : Inter-Item Covariance Matrix

The Box’s M Plot has been debated among data scientists about its effective use. So, to ensure that there is an equality of covariance among your dependent variables is to split the file and test the covariance across the dependent variables as shown above. As a rule of thumb, if the corresponding covariance values are not above 3 times each other’s values, the MANOVA will be quite robust.

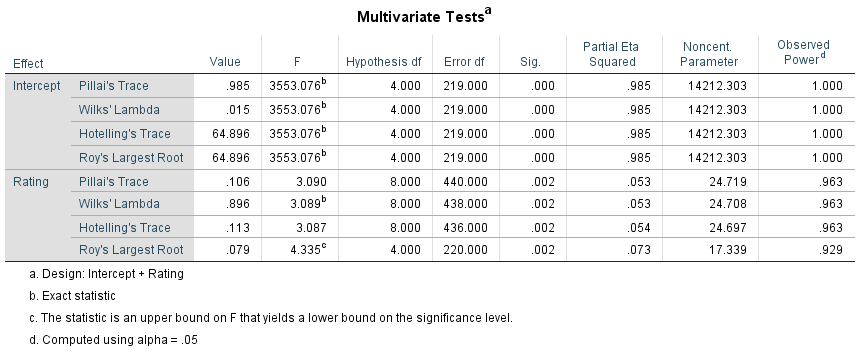


Figure : Multivariate Tests

Using Wilks’ Lambda as it is one of the most used multivariate tests in statistics, we can conclude that the p-values for the ratings are statistically significant at the 0.05 significance level because the p-value is less than .05. The observed power being at 96.3% means that there is only a 3.7% chance of a Type I Error occurring in the MANOVA. The effect size is extremely low as it sits at less than 0.1, this means that the differences in the variables is not of concern.

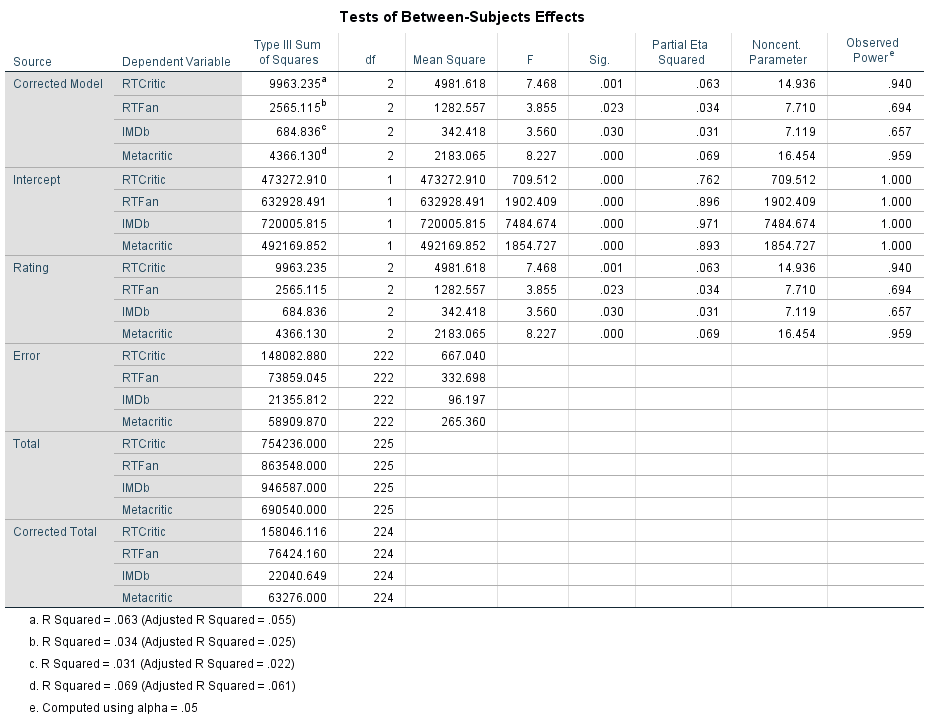


Figure : Tests of Between-Subject Effects

This table shows the univariate ANOVA’s which means that the inter-correlation between the dependent variables is void. Take the IMDb variable for example. The significance value for IMDb shows a 0.3 which indicates no significant differences among the rating groups in respect to the IMDb variable.

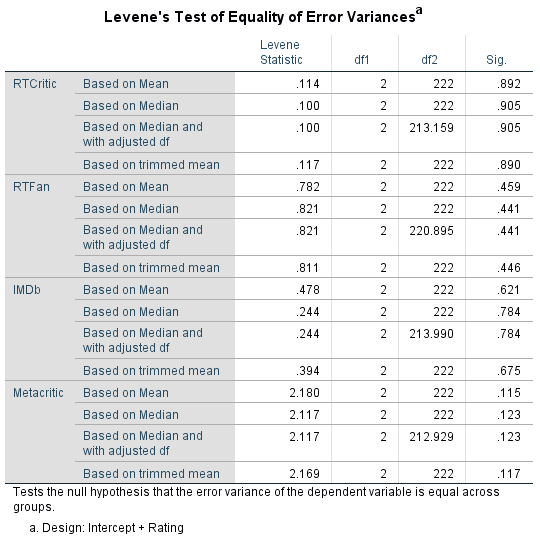
  
On this test, the p-value for the mean of each dependent variable respectively is .892, .459, .621 and .115. From this result, we can say that the assumption of equal variances is met as the p-value is greater than .05. which gives confidence that we can trust the univariate test results.

Figure : Levene's Test of Equality of Error Variance

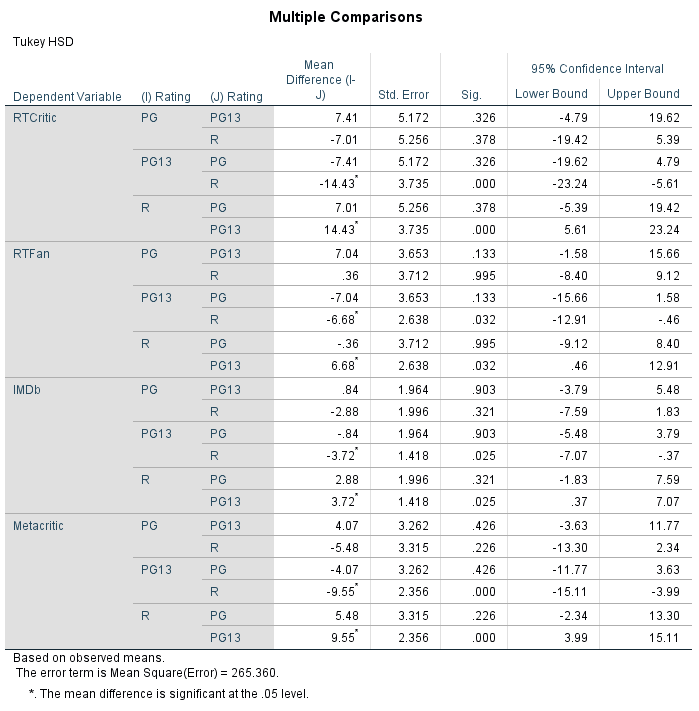


Figure : Post-Hoc Test

This Post-Hoc Test measures the groups in the independent variable against each other in relation to the dependent variables, highlighting the mean difference and significance values. From the p value, we can see that in most cases there is no significant difference as most p-values are greater than 0.05. As a brief example to understand the mean difference, take RTCritic and their first rating of PG compared to PG13. We can see a mean difference of 7.41 which means that on average RTCritic have given a 7.4% better rating to movies rated PG than they have to movies rated PG13.

As the p-value is greater than 0.5, we fail to reject the null hypothesis, giving us the answer that there are differences in the critics’ ratings of movies depending on the age rating of the movie.

**MULTI-LEVEL REGRESSION**

The second dataset I have chosen for this assignment to run a multi-level regression on is a dataset on Car Prices. This dataset contains over 200 rows with over 20 independent variables to access to run this multi-level model. The data has been sourced from Kaggle.com (Kumar, 2019). The main independent variables focus on the car’s design and features such as car length, horsepower, wheelbase, engine type, car body, car drive, car brand, engine size, fuel type and other variables which can factor in determining the price of a car.

Firstly, to assess the assumptions of linearity and homoscedasticity of each independent variable against the dependent variable (price), I created scatter plots in SPSS.

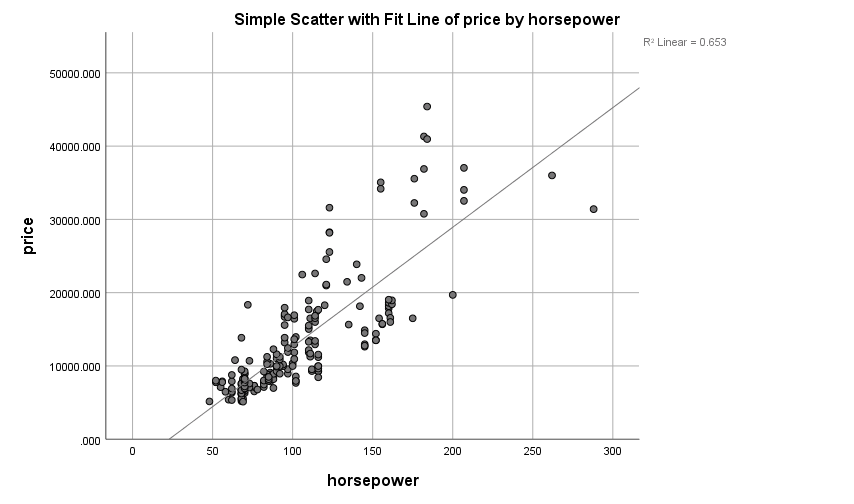


Figure 8: Scatter Plot with Fit Line of Price vs. Horsepower

This scatter plot shows a decent fitted line with an R-squared value of 0.653, this indicates that horsepower can explain 65% of all variability of the response data around the mean.

The R-Squared value measures how close the data is to a fitted line for regression.

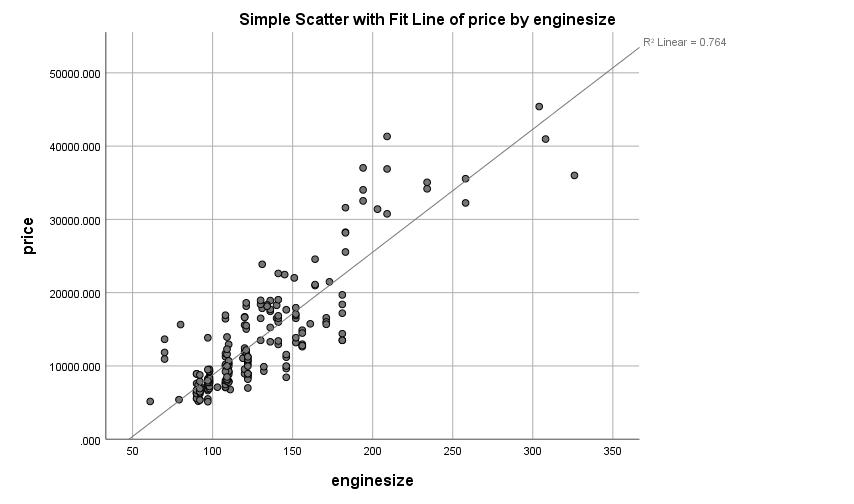


Figure 9: Scatter Plot with Fit Line of Price vs. Enginesize

This scatter plot highlights the linearity of Price measured against Engine Size. As shown from the modelled fitted line, there is high homoscedasticity with an R-Squared value of 76%, which accounts for over ¾ of all variability of the response data around the mean.



Figure 10: Scatter Plot with Fit Line of Price vs. Wheelbase

From this scatter plot, we can see the fitted line is not as strongly correlated and the variables pitted against each other do not match close enough to the fitted line to include in the regression model. The R-Squared value is 0.334 which means that only 33% of the variability is accounted for.

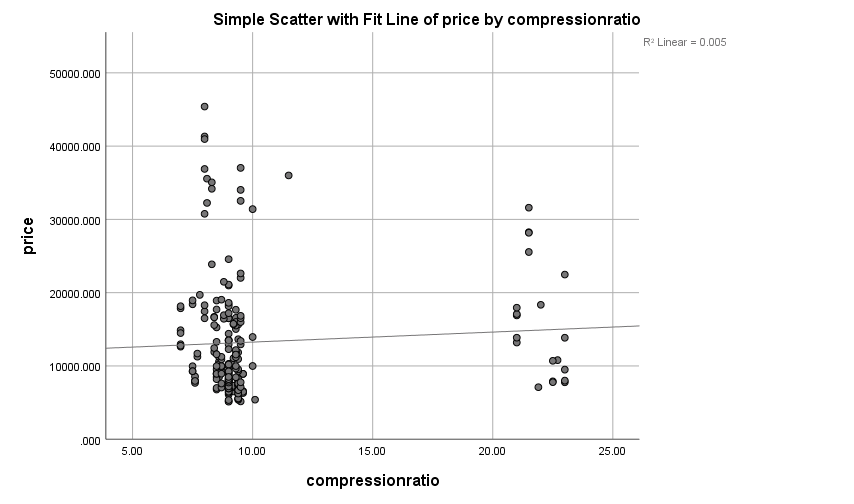


Figure 11: Scatter Plot with Fit Line of Price vs. Compression Ratio

This scatter plot shows that the fitted line is not strongly correlated at all and that the variables do not match to the fitted line to include in the regression model. The R-Squared value is 0.005 which means that less than 1% of the variability is accounted for. This variable is rendered useless and cannot be used in the regression model.

So, for my regression analysis, my Dependent Variable (DV) is price, and my Independent Variables (IV) are going to be engine size and horsepower. It is good to only have a small number of IV’s as a larger number can increase the chance of an error and it will reduce the statistical power of the test.

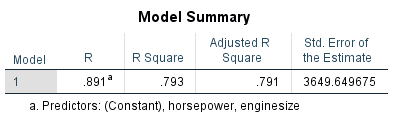


Figure 12: Model Summary of Regression

The R-Squared value is .793. This rounds down to 79% which shows that Horsepower and Engine Size account for 79% of the variance of the price of cars.

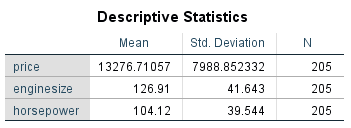


Figure 13: Descriptive statistics for the variables involved in the multiple regression model

This is the descriptive statistics associated with each variable disclosed in the multiple regression model

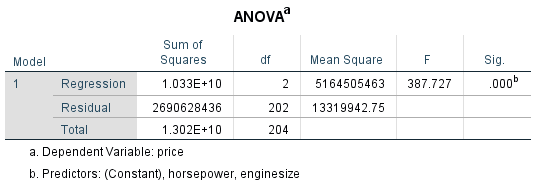


Figure 14: ANOVA for the regression model

As shown, our p-value (Sig.) < 0.05 which indicates that the regression statistically significant. This means that R-Squared > 0. This means my predictors (IVs) count for a significant amount of variance in the price of cars. F (2, 202) = 387.73, p < .001, R-Squared = .79.

The F-Test assesses multiple coefficients simultaneously. It compares the fits of different linear models for regression.

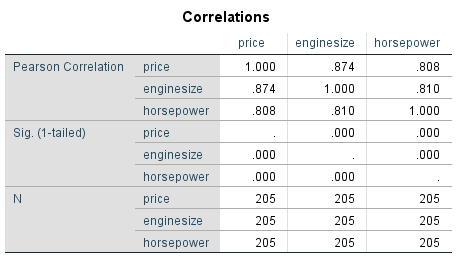


Figure 15: Correlations of the variables

The correlation table is essentially what the multiple regression is based upon. We can see, using the Pearson-Correlation that each variable is highly correlated to each other with over .80 assigned to each correlation.

In a real-life point of view, it is safe to assume that the higher the horsepower on a vehicle, the bigger and more expensive the engine which overall generates a higher price for the car. Also, the bigger the engine on a car, the higher the horsepower which in turn sets the price higher for a car.

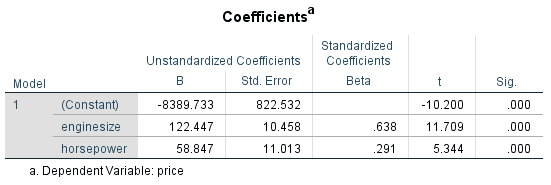


Figure 16: Coefficient Table for Multiple Regression

The coefficients table looks at each predictor value (IV) individually instead of taken as part of a group. This tells us if the IV was statistically significant on its own right.

If we take the p-value at .05, we see that each IV has a significance value 0f .000 which shows that the p-value < .05 which shows each IV is statistically significant.

Enginesize Sig. (p < .001)

Horsepower Sig. (p < .001)

If a test is significant, this means that the amount of unique variance that the predictor accounts for is statistically significant.

Unique variance means that enginesize explains something in the price that horsepower has not and vice versa.

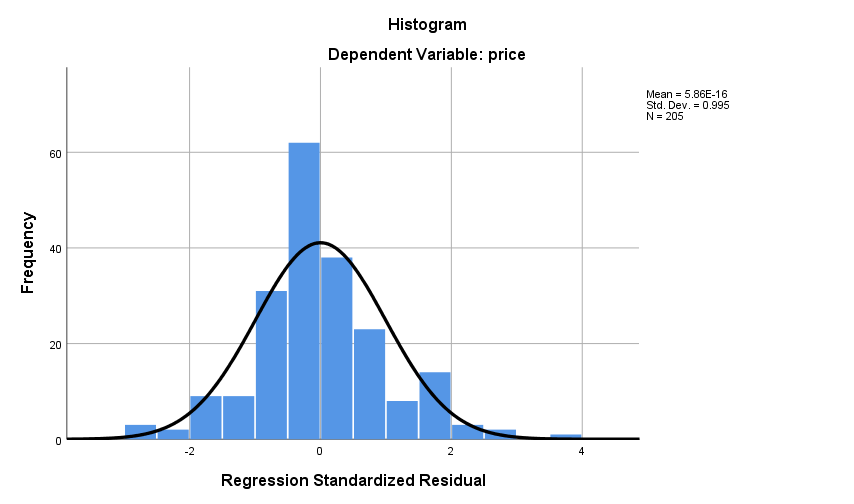


Figure 17: Histogram of Regression Standardised Residuals

From the graph, we can see it is normally distributed. The residuals are what is left over after being predicted by the independent variables as a multiple regression equation.

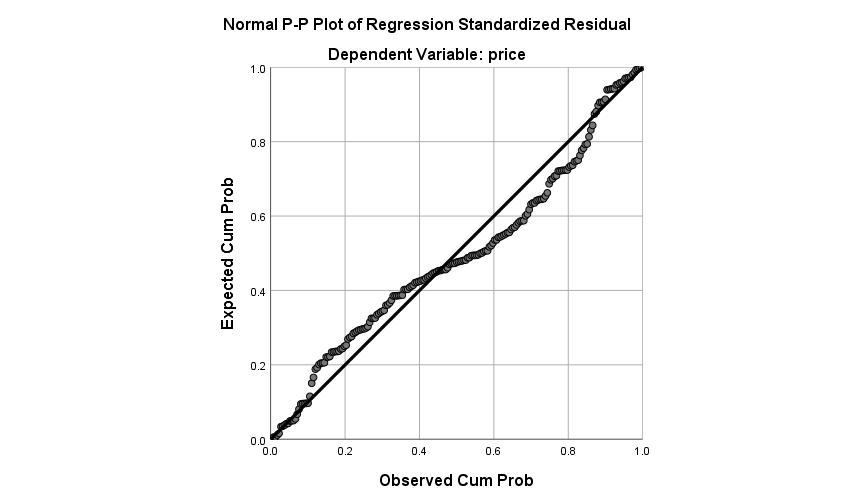


Figure 18: P-P Plot of Expected and Observed Probability for Regression Standardised Residuals

It is a good sign that all plots are located extremely close the line of best fit.

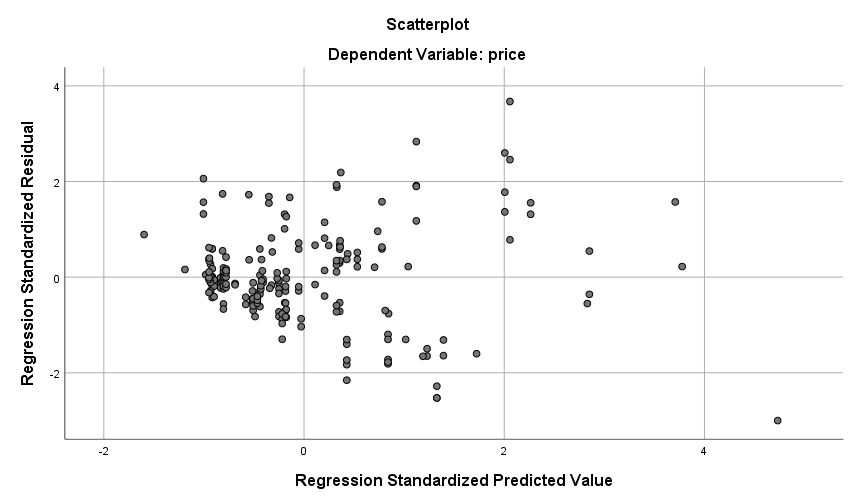


Figure 19: Scatterplot of Regression Standardised Predicted Value (x-axis) and Regression Standardised Residual (y-axis)

From this graph, we can see a lot of the plots are bunched up on the left side of the graph which shows heteroscedasticity and multi-collinearity.

**EFFECTIVE SIZE**

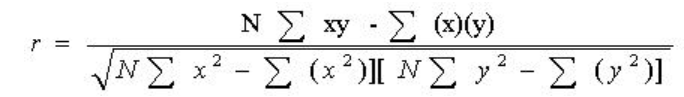
Effect size is an easy way of comprehending the contrasts between two separate groups. Effect size has the upper hand over testing of statistical significance exclusively. Effect size brings to light the major contrasts rather than confounding this with sample size. (Coe, 2002)

Effect size exists to highlight contrasts in correlations and ANOVAs. They are essential when attempting to find the correct sample size. The smaller the effect size means the difference is unimportant while a larger scaled effect size means the difference is of importance.

To highlight this further if we have a dataset based on men and women’s height. It is common knowledge that men have a taller average height in comparison to women. The difference between men’s height and women’s height is the effect size. The bigger the effect size is, the greater the gap between the height of men compared to women. Statistical effect size aids the analyst to conclude if the difference is conclusive or if it is due to a change of factors.

To delve further into the statistical area of Effective Size, there are a plethora of types of effect sizes to use on statistical tests.

Firstly, there is Pearson’s r correlation, this was created by Karl Pearson and is the most popular when performing statistical tests. The parameter of the effective size is denoted by the sign ‘r’. The value of this denotation ranges from -1 to +1. Generally, the effect size is low if r sits around 0.1, 0.3 is average and if r is more than 0.5, it is large. The formula used to compute the Pearson Correlation is as follows:

(Statistics Solutions)

Denoting symbols:

r = correlation coefficient

N = number of pairs of scores

Σxy = sum of the products of paired scores

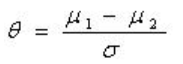
Σx = sum of x scores

Σy = sum of y scores

Σx2 = sum of squared x scores

Σy2 = sum of squared y scores

When the mean of a population and standard deviation is the base of the statistical test, then the standardized means difference method is used to produce the effect size. The following formula is used when applying this method:



The effect size is found by taking the two populations mean differences and dividing them by the standard deviation.

There is a wide variety of tests to be sued to identify the effect size, each test depends on the statistical analysis being done on which datasets and what category of data is being used.

**STATISTICAL POWER**

As Effect Size has been discussed above, statistical power is also of very important use in statistics. Statistical Power is a prime function of the three following factors: Sample size, Effect size, Level of Significance (P). There is a secondary factor. When 2/3 primary factors are known, you can calculate the third. Also, when the statistical power and only one primary factor is known, you sample size can be calculated which, in turn, gives you the level of significance.

When a study is being carried out, statistical power is the likelihood of getting statistical significance when you reject the null hypothesis. However, studies may show no significance. This can be because there really is no true significance or because the study failed to detect the significance giving a false negative result. False negative results come from a poorly designed study or because the study was not big enough. Poor design could be because of inaccurate measurements or a sample size being too small. (Cohen, 2013)

There are two types of errors to be careful of: Type I and Type II Errors.

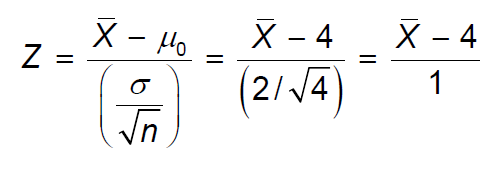
Type I Errors happen if the null hypothesis is rejected to accept the alternative hypothesis when the null hypothesis should have not been rejected.

Type II Errors happen if the null hypothesis is failed to be rejected when it should have been rejected and the alternative hypothesis should have been accepted in its place.

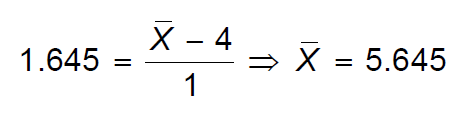
The chance of making a Type I Error is 5% with a 95% success rate however when multiple tests are being conducted, the possibility of a Type I Error being made compounds for each statistical test being done.

However, there is a way to combat the compounding of Type I Errors. By using the Bonferroni Correction, you can keep the Type I Error probability at 5% by dividing the probability by the number of tests being ran to give you a smaller criterion for significance. However, this does lead to a loss of statistical power. The power shows the ability of a test to reject or fail to reject the null hypothesis correctly. A high power gives a lower probability, and a lower power gives a higher probability.

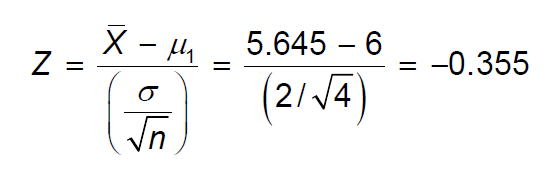
An acceptable level of power is 80% or above. Let us say the average time of exercise per week is 4 hours. We set this as the null hypothesis, and we set the alternative hypothesis at 6 hours a week with a standard deviation of 2 hours. To calculate this result, we use this formula:



At 5% significance level, the constraint for the test is to reject the null hypothesis if Z > 1.645. These are the results of the above formula.



We then take that 5.645 to calculate the Z statistic using the alternative hypothesis which gives us the solution:



P(Z > -0.355) = 0.6368

From these results, we round up to 64% which means the statistical power of this test is estimated to be 64%. As 64 < 80, we can safely say that this test ‘lacks power’ due to the small sample size. So, to increase the power of the test you can:

Increase the effect size to be detected, decrease the variability in the sample or increase the significance level of the test.

**References**

Allen, A. (2017) *Comparing Numerical Movie Review Scores*, *Kaggle.com*. Available at: https://kaggle.com/antallen/comparing-numerical-movie-review-scores (Accessed: 9 January 2021).

Coe, R. (2002) *It’s the effect size, stupid: what effect size is and why it is important*. Education-line. Available at: https://www.leeds.ac.uk/educol/documents/00002182.htm (Accessed: 7 January 2021).

Cohen, J. (2013) *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Kumar, M. (no date) *Car Price Prediction Multiple Linear Regression*, *Kaggle.com*. Available at: https://kaggle.com/hellbuoy/car-price-prediction (Accessed: 9 January 2021).

Statistics Solutions (no date) ‘Effect Size’, *Statistics Solutions*. Available at: https://www.statisticssolutions.com/statistical-analyses-effect-size/ (Accessed: 7 January 2021).