BlogPost #10: STATISTICAL POWER, TYPE I/II ERRORS, EFFECT SIZE

Having previously scoped out a research methodology, I am now at the stage of this project where I need to focus carefully on the research methods for data collection and statistical analysis so that a response to my hypotheses makes sense in statistical terms. I have been reflecting on the nature of the data collection processes that I will devise and deploy, and the 1st draft of the research e-Questionnaire that will be the mechanism for acquiring the information required has been designed and created. (This development process, together with a discussion on the questionnaire design will be discussed in a subsequent post shortly).

As I am thinking more about the data processing part of the project still to come, it is increasingly clear that the data collection and statistical analysis that I devised and worked through in the pilot study MSc dissertation was naive and lacking in the level of robust statistical methodology planning that is now essential for this PhD project.

For example: in the e-QNR that I devised in the pilot study, although I believe that the rationale was sound – that is, identifying 5 psycho-educational dimensions or factors to explore (Learning Related Emotions; Anxiety, Regulation & Motivation; Self Esteem; Self Efficacy; Learned Helplessness) which collectively were graphically represented to create a profile corresponding to each respondent's answers – a lack of understanding about some of the fundamental statistical analysis processes beyond the application of simple tests of significance, quite possibly rendered the conclusions derived from the output of the statistical tests conducted hopelessly inaccurate. So for this current research a much more coherent understanding of stats principles as applied to data collection and analysis is required.

Despite having a mathematical background, albeit not with a statistics focus, the notions of 'effect size' and 'statistical power' are new to me and given an increasing shift of focus away from merely reporting 'p-values' and making conclusions based on levels of significance in research papers accepted for publication by peer-reviewed journals, demonstrating expertise in wider processes of stats analysis will contribute to results that are more robust. At least I am familiar with the concepts of Type I and Type II errors so let us start with a simple summary of these:

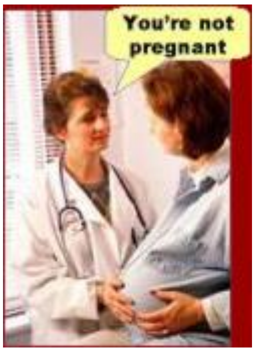
Type I and Type II errors

A Type I error occurs when the data analysis produces a result that is a **FALSE POSITIVE**.

For example: A gentleman who is concerned about his health visits his doctor. He tells the doctor that recently he has been feeling nauseous early every morning, that he has been told by close friends that he is showing increasingly obvious mood swings and that he is developing a

curious pre-disposition for odd combinations of food – his favorite at the time being strawberry jam and cucumber sandwiches to which the doctor remarked that these probably hadn't helped his obvious weight problem. His doctor considers these symptoms carefully and despite his better judgment, decides that he needs to administer a pregnancy test because the gentleman is exhibiting typical characteristics of the condition. After doing so, incredibly the test indicates that the gentleman is pregnant = **false positive**.

So incorrectly rejecting the null hypothesis is called a Type I error. If we reject the null hypothesis when $p < 5\%$, this means that we would reject the null hypothesis in 5% of cases when it is, in fact, TRUE.



Conversely, a Type II error occurs when the data analysis produces a result that is a **FALSE NEGATIVE.**

Oddly enough, the gentleman's wife visits the same doctor the following day. She is clearly in some discomfort as she waddles, puffing and blowing, into the doctor's consulting room. She is supporting the small of her back with her hand and presents a highly distended abdomen, so much so that her clothing is so tightly strained it is set to burst. As she sits down carefully with a sigh of relief, the doctor exclaims how surprised he is at the difference in her physical appearance compared with her last visit (picking up from his notes that in an earlier consultation some 8 months previously, she had been a slim, picture of fine health). Not wanting to appear rude by commenting unduly on her changed appearance, the doctor asks her why she has visited his surgery that day. Curiously, the lady describes almost identical symptoms to those mentioned to the doctor by her husband the previous day (except the sandwiches – she'd taken a recent fancy to prunes with goats' cheese) so the doctor decides again to administer a pregnancy test and after doing so, much to his surprise, the test shows no indication of a pregnancy = **false negative.**

So failing to reject the null hypothesis when it should have been rejected is a Type II error. If the probability of making a Type II error is ß, then 1 – ß is the probability of rejecting the null hypothesis when it SHOULD be rejected – that is, a CORRECT conclusion. **This is known as the POWER of the test** – more below;

Effect size

Effect size is a concept that appears to be a relatively new idea in the realm of statistical analyses and challenges the traditional approach that the **p-value** is the most important outcome that directs the researcher's response to the research hypotheses.

Effect size values are a measure of either the magnitude of associations or the magnitude of differences, depending on the nature of the data sets being analysed. In fact, for associations, the most frequently used measures to determine the strength (magnitude = 'size') of association are correlation or regression coefficients.

We know that the p-value outcome is an indication of **statistical significance**, that is, the probability of whether an outcome has naturally occurred as a result of chance, or is otherwise is indicating that something has happened, that is there has been an 'effect' and the outcome observed has **not** occurred randomly by chance.

Statistical tests that generate this measure of significance are widely used by researchers and although any 'level of significance' could be used as the determining cut-off point, we all know that $p=0.05$ is traditionally used as the borderline probability level where a value of $p < 0.05$ leads researchers to conclude that they have a 'significant result' and that therefore this result has NOT occurred by chance.

However, my literature trawl on this topic is aiding my understanding of the lack of oommph that exists when relying on just the p-value alone to come to a conclusion. Sullivan (2012), amongst others, tells us that when a sufficiently large sample size is employed, a test of significance used to determine whether there has been an effect or not (her context is in medical research) will almost certainly 'demonstrate a significant difference, unless there is no effect whatsoever'. For example, in a large sample of say, $n=20,000$ that is being used to explore the effect of a drug intervention to mediate a medical condition, due to the size of the sample a statistical test will almost invariably determine that there is a significant between-groups difference in the (mean) effect of the drug even though that actual (or absolute) difference (between the groups' means) is very small. We learn from this that whereas significance tests are influenced by sample size, effect size is not because it is an absolute measure, usually of this difference between means.

Effect size is easy to calculate, indeed the simplest result is merely the absolute difference between the means of two independent groups' data sets. An improved measure is derived by dividing this result by the standard deviation of either group and in this form, the effect size is referred to as ‘d’, more usually ‘Cohen’s ‘d” after the originator of the idea (Cohen, 1988). There are various other measures for effect size usefully collected into a summary paper by Thalheimer (2002).

Cohen labelled, rather arbitrarily it seems, the magnitudes of effect size measures at d=0.2 (small), d=0.5 (medium) d=0.8 (large) and d=1.3 (very large) and although these labels don’t appear to account for the impact that other factors may have on the variables such as the accuracy of the data gathering tool or the diversity of the study’s background population but despite this, it appears that these labels or designations are widely used by researchers, and hence their meanings are commonly understood.

Sullivan (2012) also very concisely summarizes what effect size means when used as a measure of the between-groups difference between means given that the data distributions of each of the groups is normal – pretty much always the case in a high proportion of research not the least due to the Central Limit Theorem. Sullivan usefully describes effect size as the ‘amount of overlap between the distributions’ and provides an easy to understand example by first reminding us that an effect size of 0 (zero) would of course indicate that there is no difference between the means of the groups and that there would be *total eclipse* of one (standardized) normal distribution over the other (my analogy). Conversely, an effect size of say, 0.7, would be indicating that the mean of group 2 is at the 69th percentile of group 1 and hence that someone from group 2 with an average (mean) score would have a higher score than 69% of the people from group 1. An interesting and very visual interpretation of this idea is presented by Magnusson (see note 3 below) where it is possible to slide one normal distribution across another to reveal the effect size v overlap relationship.

Statistical Power

A concise paper by Skrivanek (2009) defines the power of a statistical test to be the probability that the null hypothesis will be rejected when it is actually false – which represents a correct decision. In contrast, the signicance level of a test provides a probability that the null hypothesis will be rejected when it is true, which is an incorrect conclusion. It is important to note that the *p*-value is a measure of the **strength of evidence** and is not, directly, a measure of the **size** or **magnitude** of the difference (between means). It is possible to have plenty of evidence for a tiny and uninteresting difference, **especially in large samples** – which is another way of saying that in a large simple, it is quite likely that we might end up with a significant difference.

So the POWER of a test is a measure of its ability to *correctly* reject the null hypothesis and where this is useful is that to be able to calculate the POWER of a test *before* the data is collected and the stats analysis is conducted, will ensure that the sample size is large enough for the purpose of the test and conversely not so large as to most likely arrive at a significant result anyway.

This is quite a difficult idea to grasp but is clearly important since it dismisses the intuitive assumption that the more data you have, the more reliable the results of the analysis will be.

It appears that amongst statisticians, a test that has 80% power or better is considered to be 'statistically powerful' so in working backwards from this, it is possible to calculate an ideal sample size that will generate this level of power given that other parameters about the data distribution are either known or can be reliably estimated. In addition, larger differences between means are obviously easier to detect and this will have a (beneficial) impact on the power of a test – that is, increasing the **effect size**. Sullivan (2009) comments on the balance between effect size and sample size by remarking that it will obviously be easier to detect a larger effect size in a small sample than if the effect size – that is the difference between the means – is small and that conversely, a smaller effect size would require a larger sample in order to correctly identify it.

So what is key in this discussion, is that finding a way to establish a sample size that is appropriate for the desired power level is very important and Sullivan guides us about how to do this by suggesting that we can either use data from a pilot study, consider the results from similar studies published by other researchers, or think carefully about the minimum difference (between means) that might be considered as important.

Refer to Skrivanek's paper for an example of the steps involved in calculating the power of a test.

Notes:

1. A further, excellent summary of the concept of Effect Size and its relationship with Null Hypothesis Testing from the University of Bath's Department of Psychology is here;
2. Another useful summary paper of the various Effect Size measures and how to calculate them is here, published by Thalheimer, W, (2002) available at: www.work-learning.com;
3. An interesting visualization of the relationship between the value of Cohen's 'd' for effect size and the corresponding overlap of two groups' normal distributions is here;

4. Paul Ellis from Hong Kong Polytechnic University provides a useful suite of webpages dedicated to Effect Size here;

References

Sullivan, G.M., Feinn, R. (2012) *Using Effect Size—or Why the P Value Is Not Enough.* Journal of Graduate Medical Education: September 2012, Vol. 4, No. 3, pp. 279-282.

Cohen, J. (1988), Statistical Power Analysis for the Behavioral Sciences, 2nd Edition. Hillsdale: Lawrence Erlbaum.

Ellis, P.D. (2009), "Thresholds for interpreting effect sizes," Available at: http://www.polyu.edu.hk/mm/effectsizefaqs/thresholds_for_interpreting_effect_sizes2.html, accessed on: 3rd July 2015.

Skrivanek, S., (2009), *Power of a Statistical Test,* Available at: http://www.morestream.com/whitepapers/download/power-stat-test.pdf, Accessed on: 30th June 2015.