A reverse look at p-values

Paul White,

Applied Statistics Group,

Faculty of Environment and Technology,

Univeristy of the West of England, Brsitol,

Bristol BS16 1QY, UK

paul.white@uwe.ac.uk

Paul Redford,

Department of Health and Social Sciences

Faculty of Health and Applied Sciences,

University of the West of England, Brsitol,

Bristol BS16 1QY, UK

paul.redford@uwe.ac.uk

James Macdonald,

Department of Health and Social Sciences

Faculty of Health and Applied Sciences,

University of the West of England, Brsitol,

Bristol BS16 1QY, UK

james.macdonald@uwe.ac.uk

***Abstract*—** An overview of *p*-values is given. The usual way of being introduced to *p*-values is by considering the rigorous development of statistical tests, then their use for decision making. This note takes a reverse view by firstly considering their use in decision making, then their distributions, before looking at the statistical test.

Keywords*— *p-values,

I. Introduction

The results of a statistical test are often summarised by a *p*-value (sometimes called a significance value). A common question when first using quantitative research methods and statistics is the question "*What exactly is a p-value*?" and "*why is p < 0.05 (often) taken to mean statistically significant?*", and "*what do we mean by significant?*" This brief note will give an introductory answer to these questions.

By way of example, the note will refer to three examples taken from this series. The three examples are given in [1, 2, 3] and it might be instructive to read through these examples to get the most out of this brief note. However, as a brief recap, these three papers

(a) Test a null hypothesis of no association between breast injury and breast cancer. The null hypothesis is rejected with $p < .001$, using the chi-square test of association. See [1].

(b) Test a null hypothesis of homogeneity of mean weight loss between those on a regular diet and those on a new diet. The null hypothesis is rejected with $p = 0.003$, using the independent samples t-test. See [2].

(c) Test of a null hypothesis of homogeneity of means when comparing reaction time between relatively old and relatively young people. In this example there is a failure to reject the null hypothesis with $p = 0.096$, using the separate variances t-test. See [2].

(d) Test a null hypothesis whether mean number of aggressive acts differ before and after exposure to violent media. In this example the null hypothesis is rejected with $p = 0.09$, using the paired samples t-test. See [3]

To motivate matters we will start with a working definition of a *p*-value (and one which, for now, deliberately avoids the word "probability"). The proposed definition is "*For a given data set and given test statistic, the p-value is the largest significance level for which there is failure to reject the null hypothesis*."

Let's break this down.

II. A first definition

For a given data set and given test statistic, the p-value is the largest significance level for which there is failure to reject the null hypothesis."

"*For a given data set*". It stands to reason that if the data were to change then the *p*-value would also change. That is, if the inputs to a test change then the outputs change.

"*For a given test statistic*". In example (b), alluded to in the introduction, the data was analysed using the independent samples t-test. It would not have been entirely unreasonable to have analysed the data using the separate variances version of the t-test (i.e. Welch's test), or perhaps, with a slight change of null hypothesis, to have analysed the data using the Mann Whitney Wilcoxon test. Likewise, in example (d), rather than using the paired samples t-test an alternative might have been to use the Wilcoxon Signed Rank test. *Different test statistics applied to the same data could give different p-values. Hence the observed or reported p-value depends on the choice of statistic.* Of course, which is the best statistic to use is dictated by the circumstances.

Contemporary practice is to reject a null hypothesis if the *p*-value is less than 0.05 (i.e. less than 1 in 20). In this sense 0.05 or 5% is a nominal significance level and is traditionally denoted by alpha. There may be situations

where testing is done at different levels; perhaps at the 10% level (i.e. alpha = 0.1), or the 1% level (i.e. alpha = 0.01).

If working at the usual 5% level then any *p*-value less than 0.05 would indicate the rejection of the null hypothesis and a claim of significance at the 5% level. If working at the 5% level and a *p*-value is greater than or equal to 0.05 then this would indicate a failure to reject the null hypothesis at the 5% level.

If working at the 10% level then any *p*-value less than 0.10 would indicate the rejection of the null hypothesis and a claim of significance at the 10% level. If working at the 10% level and a *p*-value is greater than or equal to 0.10 then this would indicate a failure to reject the null hypothesis at the 10% level.

If working at the 1% level then any *p*-value less than 0.01 would indicate the rejection of the null hypothesis and a claim of significance at the 1% level. If working at the 1% level and a *p*-value is greater or equal to 0.01 then this would indicate a failure to reject the null hypothesis at the 1% level.

So, for instance, if an analysis gave a *p*-value equal to 0.09 then there would be a rejection of the null hypothesis if working to a pre-declared and reasoned 0.10 significance level but there would be failure to reject the null hypothesis if working to the usual 0.05 level.

Now suppose, quite bizarrely that a researcher was going to work at the nominal 9% significance level (i.e. alpha = 0.09) and had a *p*-value exactly equal to 0.09. Would we reject the null hypothesis in this situation? No! we are right on the tipping point and that is why the first definition of a *p*-value (deliberately avoiding the word "probability") contained the wording "*the p-value is the largest significance level for which there is failure to reject the null hypothesis.*"

So, there we have it; we have a definition of a *p*-value of "*For a given data set and given test statistic, the p-value is the largest significance level for which there is failure to reject the null hypothesis*." The problem with this definition is it simply says how the *p*-value might change (e.g. different data, different test statistic) and how to make a statistical decision. It does not give a great insight into what the *p*-value tries to summarise. For this we need to look a little further into hypothesis testing and how statistical tests are constructed.

.

III. WHAT DO P-VALUES LOOK LIKE?

Let's do a mind experiment. Suppose we research all of the mathematical and statistical assumptions which underpin the independent samples t-test. Further suppose we generate data to meet these assumptions (e.g. we generate two independent random samples from the same normal distribution) and for this sample we calculate the *t*-statistic and the *p*-value.

In this hypothetical situation the null hypothesis is true because we have generated data from the same normal distribution and any difference in the two *sample* means can be ascribed to chance natural variation arising from random sampling.

Now let's repeat the above process and calculate a second *p*-value. In fact, let's go through this procedure one million times. We would now have 1,000,000 *p*-values all generated under perfect conditions and all generated under a true null hypothesis. Suppose we create a histogram of these 1,000,000 *p*-values. What would the histogram look like?

For fun we have done this. The histogram of the 1,000,000 *p*-values is given in Figure 1.

The histogram in Figure 1 looks to be a uniform distribution whereby all values between 0 and 1 are equally likely. Put another way, 1% of the time the *p*-values are smaller than 0.01; 2% of the time the *p*-values are smaller than 0.02; 5% of the time the *p*-values are smaller than 0.05; 10% of the time the *p*-values are smaller than 0.1; 30% of the time the *p*-values are smaller than 0.30 and so on. In general, *X*% of the *p*-values are smaller than *X*/100.

When mathematical statisticians design statistical tests, they design them so that *if* the null hypothesis is true, *if* assumptions are satisfied and *if* the correct statistical test is used, then the resulting *p*-values will be uniformly distributed between [0, 1]. Sometimes, a mathematical statistician, in the absence of being able to develop a precise test, will develop an approximate test and the *p*-values under these approximate tests have a distribution which is approximately uniformly distribution between [0, 1].

It is worth restating this: in an idealized world, if the null hypothesis is true, and if assumptions are satisfied, and if the most appropriate test statistic is used, then the resulting *p*-value is a random instance from the uniform distribution with support [0, 1]. This is true irrespective of sample size.

Of course, in any practical situation, you will only have one *p*-value from your test.

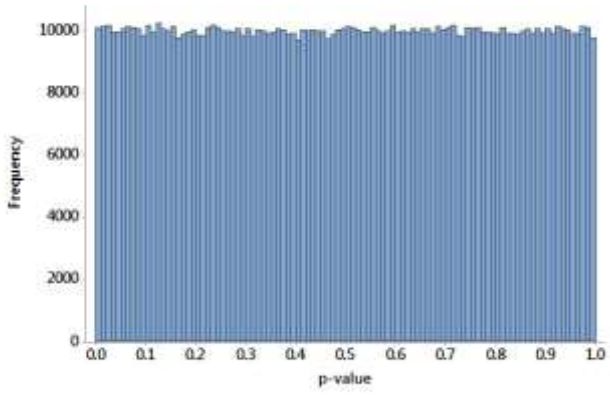

Figure 1 1,000,000 p-values generated when the null hypothesis is true

IV. What do p-values look like when the Null hypothesis is fasle?

We could now do a second mind experiment. We could consider what the distribution of *p*-values would look like when the null hypothesis is false (i.e. when we should be rejecting the null hypothesis). This is a bit deceptive. There is only one way in which the null hypothesis can be true (e.g. identical means) but there are infinitely many ways in which the alternative hypothesis can be true e.g. means differing by 1, or by 2, or 2.3, or 10. Not only that but when the alternative hypothesis is true the distribution of *p*-values would also depend on sample sizes. Suffice to say, Figure 2 is an example histogram of 100,000 *p*-values generated in a particular instance of when the null hypothesis is false (alternative hypothesis true). To extend this, we have produced a histogram of 100,000 *p*-values when means differ by +1 and another histogram of 100,000 *p*-values when means differ by +1.5 (we will spare you the nitty-gritty details). Without looking at the graphics to follow: "*What would you imagine the two histograms to look like*?"

The two histograms alluded to are given in Figure 2 and Figure 3.

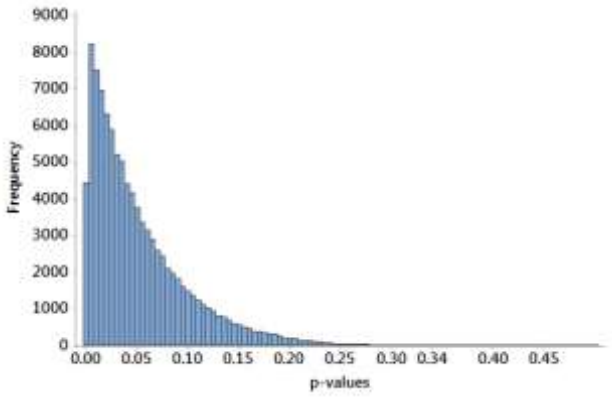

Figure 2 Histogram of 100,000 *p*-values in a particular instance when the null hypothesis is false

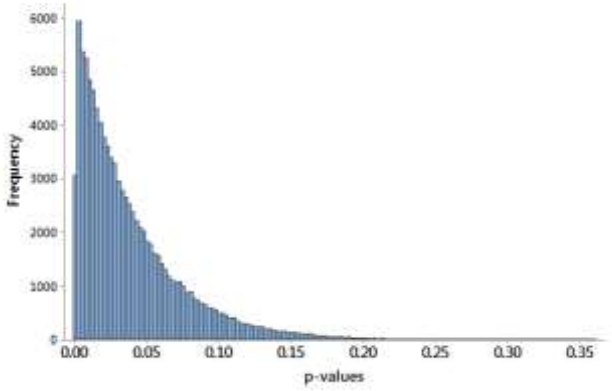

Figure 3 Histogram of 100,000 *p*-values in a particular instance when null hypothesis false

In both Figure 2 and Figure 3 the distribution of the *p*-values are no longer uniform; they are both positively skewed with much smaller mean values than those shown in Figure 1. It is noticeable that when the effect size increases (and all else remains the same) that the distribution of *p*-values has a greater cluster closer to zero (Figure 2 is for a mean difference of +1, Figure 3 is for a mean difference of +1.5).

Figure 2 and Figure 3 display a general feature of statistical tests; the mathematical statistician designs the test to have small (low) values for a *p*-value when the null hypothesis is false. A small *p*-value leads to the rejection of null hypothesis.

V. p-values and Probability

In the Introduction, reference was made to four examples. The first example was concerned with the association between breast cancer and breast injury. These data were analysed using the chi-square test of association and the calculated value of the chi-square statistic was $\chi^2 = 34.388$. This returned a p-value of $p < 0.001$. What does this mean? *Suppose the null hypothesis is true* and we ask the question "what is the probability of getting a chi-square value of 34.388 or larger assuming the null hypothesis is true". This probability is the *p*-value; this probability is less than 0.001.

The second example was concerned with whether mean weight loss between those on a regular diet differed from those on a new diet. These data were analysed using the independent samples t-test and the absolute value of the *t*-statistic was $t = 3.078$. This returned a *p*-value of $p = 0.003$. What does this mean? Suppose the null hypothesis is true and we ask the question "what is the probability of getting a *t*-statistic with absolute value of 3.078 or larger assuming the null hypothesis is true?". This probability is the *p*-value; this probability is 0.003. Likewise, the third example was concerned whether the mean reaction times differed between relatively older people and relatively younger people. These data were analysed using the separate variances t-test and the absolute value of the *t*-statistic was $t = 1.746$. This returned a *p*-value of $p = 0.096$. What does this mean? We can ask the same question again and base it on the assumption that the null hypothesis is true. The answer to this question is the *p*-value; this probability is 0.096.

The fourth example was concerned whether the mean aggressive behaviour differed pre- and post- intervention. These data were analysed using the paired samples t-test and the absolute value of the *t*-statistic was $t = 3.343$. This returned a *p*-value of $p = 0.009$. What does this mean? Suppose the null hypothesis is true and we ask the question, what is the probability of getting a t-statistic with absolute

value of 3.343 or larger, assuming the null hypothesis is true. This probability is the *p*-value; this probability is 0.009.

In summary *p*-values related to a null hypothesis. They do not relate to the alternative hypothesis. The -*p*-value is predicated on a temporary assumption that the null hypothesis is true, that all underpinning assumptions hold, and relates to the test statistic used.

Importantly, *p*-values should **not** be interpreted as being a probability of the null hypothesis being true. A null hypothesis is either true or false. If a null hypothesis is true then it is true with probability 1. If a null hypothesis is false then it is false with a probability of 1.

VI. P < 0.05

The *p*-value is a summary of whether observed data deviates from a point null hypothesis by an amount which can be reasonably ascribed to chance deviations expected under random sampling. The *p*-value directly relates to the null hypothesis. Contemporary practice is to reject the null hypothesis if the observed *p*-value is less than 0.05. Where does this threshold come from?

The use of alpha = 0.05 as being "standard" appears to have gained traction from the 1920's onwards. R A Fisher (1890 – 1962) a pioneering statistician and geneticist, described as "*a genius who almost single-handedly created the foundations for modern statistical science*" [4] produced one of the first ever books on research methods and statistics. In this text [5, p 504] Fisher wrote

"*... it is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials.*"

Hence, one-in-twenty or 0.05.

It should be acknowledged that Fisher was not a stubborn advocate of the 5% level. Later in the same text, he wrote "*If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance."*

However, his book was so impactful that by the 1950s the terminology "statistically significant" was interchangeable with $p < 0.05$.

Of course, fixed point significance testing and null hypothesis testing is not without criticism e.g. see [6] but that is a story for another day.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the Learning and Teaching Initiative in the Faculty of Health and Applied Sciences, University of the West of England, Bristol in supporting the wider Qualitative and Quantitative Methods teaching programme.

SERIES BIBLIOGRAPHY

White P, Redford PC, and Macdonald J (2019) A primer on validity and design terminology in comparative designs, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –4

White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the chi-squared test of association (2 by 2), *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –4

White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the independent samples t-test and the Welch test, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –6

White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the paired samples t-test, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –5

White P, Redford PC, and Macdonald J (2019) A primer on statistical hypotheses and statistical errors, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –4

White P, Redford PC, and Macdonald J (2019) A reverse look at p-values, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –5.

White P, Redford PC, and Macdonald J (2019) That assumption of normality, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –11

White P, Redford PC, and Macdonald J (2019) Cohen's *d* for two independent samples, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –4

REFERENCES

[1] White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the chi-squared test of association (2 by 2), Qualitative and Quantative Research Methods Project, University of the West of England, 1 –4

[2] White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the independent samples t-test and the Welch test, Qualitative and Quantative Research Methods Project, University of the West of England, 1 –6

[3] White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the paired samples t-test, Qualitative and Quantative Research Methods Project, University of the West of England, 1 –6

[4] Hald, A (1998). A History of Mathematical Statistics. New York: Wiley

[5] . Fisher RA (1925), Statistical Methods for Research Workers, Oliver & Boyd (Edinburgh)

[6] Cohen J (1994), The Earth is round ($p < 0.05$), *American Psychologist*, Vol.49. No. 12, 997-1003