A primer on statistical hypotheses and statistical errors

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

paul2.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 abstract overview of some of the logic and statistical considerations in quantitative inference is given.

Keywords— research questions, research hypotheses, statistical hypotheses, statistical errors

I. Introduction

Statistical methods are frequently used in the systematic pursuit of knowledge. At a conceptual level an empirical research model is a "*question asking process*" coupled with "*an answer producing process*". Unfortunately, in practice the "right" questions are not always asked and in practice the correct answer is not always obtained even if the right question was posed. In general each new piece of research suggests further new research questions and is better seen as a "question asking", "question answering", and "question generating" process. The entire process is fraught with difficulties and challenges and it is commonly acknowledged that it is extremely difficult to undertake high quality research.

In the following, we take a high level overview of some aspects of quantitative research predominantly focusing on the logic surrounding statistical hypotheses and potential statistical errors in inference.

A usual starting point in research is to survey extant literature on the topic and all relevant articles are critically reviewed. The "how to critically read a quantitative article" is a topic in its own right and good sources to review include [1], [2], [3] and [4].

Critically reading the literature, and talking to other knowledgeable people, would undoubtedly suggest new areas of research and research questions. Some might initially take the view that research questions have to be ground breaking and extremely novel. This is not true. Undertaking research is similar to do doing a never ending jigsaw. Pieces of the jigsaw are put in place one or two pieces at a time; if a piece does not fit then it should not be forced into the puzzle as that would take the jigsaw off in the wrong direction and at some point the pieces would have be pulled out and new, correct pieces sought. It is much better to put in one small piece of the jigsaw which correctly fits.

Distinct, but related to the Research Question, are Scientific Hypotheses (or Research Hypotheses) which are testable statements whose truth or falsity can be examined by collecting relevant data. Immediately one may wonder how to test these ideas i.e. how to go about examining the truth of a testable scientific or research hypothesis. The "how to go about testing a scientific hypothesis" is known as a *design* e.g. an experimental design, or an observational or correlational design. Of course, the form of any design will be shaped by both the research question and ethical considerations. In the following we will provide an example of a simple research question and research hypotheses (see Section II), and an example (Section III) before moving on to statistical hypotheses (Section IV), and then discussing potential statistical errors (Section V).

II. Research and Scientific Hypotheses

Empiric quantitative research invariably starts with a knowledgeable person or persons (informed by extant literature) proposing a testable idea or testable question on some phenomenon of interest. A testable question is one in which we *believe* the truth or falsity of the question can be investigated by recording facts (data) on the phenomenon of interest. The testable question is usually phrased as a research hypothesis or a scientific hypothesis. *If* the question can be investigated by conducting an experiment *then* the terminology "research hypothesis" or "scientific hypothesis" is usually referred to as an "experimental hypothesis".

A scientific hypothesis (aka a research hypothesis) is a knowledgeable statement that is tentatively advanced to account for particular facts, or to support a reasoned prediction. Scientific hypotheses can usually be stated in the logical form of the general implication and are confined to questions for which the truth or falsity can be investigated by experimentation or observation. The terminology "*the*

logical form of the general implication" means that the scientific hypothesis can be expressed using an *if-then* structure e.g., "*if* aspirin is taken *then* pain will be reduced", or "*if* the test tube is heated *then* the speed of reaction will increase". Similarly, by way of example a research question might be "Does smoking affect blood pressure?" and pertinent scientific hypotheses, informed by the literature or past observation, might be stated as "*if* smoke cigarettes *then* blood pressure will increase". As an aside, and as will be discussed later, predictive hypotheses do not in general lead to one-sided statistical hypotheses or one-tailed tests.

Note that (a) researchers will probably not put their research statements directly into an *if then* format and (b) not all *if then* statements would be scientific hypotheses. On this last point, consider an assertion that headless ghosts will ride horses faster than ghost with heads (*if* headless *then* ride quicker). In this situation there is no way that data on ghosts riding horses can be obtained; hence we cannot get to the truth of the matter by collecting data.

Statistical hypotheses arise from considering the how data under a proposed design would relatively look (a) if the scientific hypothesis is true and (b) if the scientific hypothesis is not true. For development purposes we do this by way of example before returning to abstraction.

III. STATISTICAL HYPOTHESES

Suppose we re-consider the earlier research question of "Does smoking affect blood pressure?". Further suppose we have identified a working definition of a smoker (e.g. we make considerations about people who used to smoke but no longer smoke, or perhaps smoke infrequently at social occasions, or smoke a pipe but not cigarettes, or smoke cannabis). Note that these considerations concerning smokers and non-smokers might lead us to consider appropriate and carefully designed inclusion-exclusion criteria. Further, suppose the informed operational definition of blood pressure is resting systolic blood pressure.

In this scenario we might have a scientific hypothesis (S1) which could be "smoking affects systolic blood pressure" and the null version of this (S0) would be "smoking does not affect systolic blood pressure". Alternatively, S1 could be written in a predictive manner, such as "smoking increases systolic blood pressure". Either way, S1 would be the research (scientific) hypothesis.

Now suppose we consider two hypothetical populations; one population being those who would meet study inclusion criteria, and would be not excluded by the exclusion criteria, and who smoke according to the definition of being a smoker; the other hypothetical population being those who would meet study inclusion criteria, and would be not excluded by the exclusion criteria, and who do not smoke according to the definition of being a non-smoker. Conceptually, each person in each population will have a numeric value for their resting systolic blood pressure. Not everyone will have the same resting blood pressure; there will be a distribution of blood pressures for each population. Suppose we let μ_1 denote the theoretical mean systolic blood pressure for the smoking population and suppose we let μ_2 denote the theoretical mean systolic blood pressure for the non-smoking population.

Now suppose that smoking does not affect blood pressure. *If* smoking does not affect blood pressure *then* naively, we might argue that the two distributions might be equal to one another and we would tentatively suggest equality of means, i.e. $\mu_1 = \mu_2$. This is a null position (a position where an effect has been nullified) and would be referred to as a null hypothesis, H_0: $\mu_1 = \mu_2$.

On the other hand, if smoking does affect systolic blood pressure then we would anticipate distributional differences between the two populations and reason $\mu_1 \neq \mu_2$. This alternative statistical hypothesis, H_1 is the hypothesis of interest (i.e. it captures the effect that is hypothesized). Accordingly, we have two mutually exclusive and exhaustive statements

H_0: $\mu_1 = \mu_2$
H_1 : $\mu_1 \neq \mu_2$.

Logically, only one of the two statistical hypotheses can be correct.

The statistical method proceeds using a process which mimics a "proof" by contradiction (or if you prefer, a "proof" by falsification). This statistical method, tentatively requires an initial assumption that H_0: $\mu_1 = \mu_2$ is the correct statement. Data is then collected. An assessment is then made to determine whether the observed data is compatible with the null hypothesis, H_0: $\mu_1 = \mu_2$. If there is a demonstrable incompatibility between data (i.e. hard facts) and H_0 then this would lead to the rejection of H_0 ; and the rejection of the null hypothesis logically leads to the acceptance of the alternative hypothesis H_1.

If on the other hand, the collected data is not greatly incompatible with hypothesis H_0 then we would fail to reject the statistical hypothesis H_0

IV. STATISTICAL HYPOTHESES IN ABSTRACTION

The development of statistical hypotheses is a two-stage process. On the one hand we consider the likely data that we would anticipate if the scientific hypothesis is true. On the other hand we consider the likely data anticipated if the scientific hypothesis is false. These considerations will allow us to make some assertions about the distribution of data or about particular aspects of a distribution. These particular aspects of the distribution, such as the mean value, are referred to as parameters.

A statistical hypothesis is a statement concerning one or more distributions or concerning one or more parameters of a

distribution. Example parameters would be the mean, or the median, or the variance, or the range, and so on.

The thought process of going from a scientific hypothesis (a general statement) to specific outcomes or manifestations is a process of deductive logic. It should be borne in mind that for the same situation two different investigators may propose different manifestations e.g. one researcher may consider *differences* in mean values whereas another research may consider the *relationship* between two variables possibly quantified by a correlation coefficient. Irrespective, the deductive logic will invariably lead an investigator into formulating two statistical hypotheses known as the null hypothesis and the alternative hypothesis which form (but sometimes with restrictions) two mutually exclusive and exhaustive statements. By exhaustive statements we mean that the two statements cover all possibilities. The "mutually exclusive" condition means that there is no overlap between the two statements and as a consequence only one of the two statements can be true.

The null hypothesis is often one of "no difference", or "no correlation" or "no association" (these are called nil-null hypotheses) or a parameter of a distribution is equal to a specific value, or the difference between two parameters is equal to a specific value. There is a reason for this. The reason for this approach is that "no difference" or "no relationship" or other precise statements about a parameter or parameters of a distribution or distributions (possibly coupled with other assumptions such as the mathematical form of the distribution) will completely specify a state of nature that will permit a precise evaluation of the data we can expect to observe including "reasonable" worst case scenarios or limits.

The process of expressing a scientific hypothesis and its logical negation into two mutually exclusive and exhaustive statistical hypotheses is known as deductive inference (i.e. an argument from general theory to a specific outcome under that theory). The statistical hypotheses created are known as the null hypothesis H_0 and the alternative hypothesis H_1. For instance, if μ denotes the mean of a population then possible null and alternative hypotheses could be

V. *

$$H_0: \mu \leq 0$$
$$H_1: \mu > 0$$

or if μ_1 denotes the mean of one population and if μ_2 denotes the mean of a second population then possible nil-null hypotheses could be

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

A statistical test is an investigation concerning the tenability of the null hypothesis. If the null hypothesis is rejected then only the alternative remains tenable.

Rejecting the null hypothesis using contemporary criteria is what we would refer to as a "*statistically significant*" finding. A statistically significant finding is **not** necessarily a substantive finding or one of being of clinical importance or of "importance" or of "significance" as used in every day speech. "Significance" in every day speech has connotations of being "important" or "major", or having consequences. It turns out that statistical tests will uncover very small effects (if they exist) providing the sample size is sufficiently large. However *if* you do want to show that an effect is of importance *then* a minimum requirement is to show that it is a statistically significant effect. Statistical significance, loosely speaking, means that the observed statistic could not reasonably (in a probabilistic sense) have occurred as a chance outcome assuming the null hypothesis to be (perfectly) true.

Failure to reject the null hypothesis does NOT prove the null hypothesis to be true nor does it mean accept the null hypothesis i.e. failure to reject the null hypothesis does NOT mean the same as "accept the null hypothesis". The logic underpinning null hypothesis testing is analogous to cases being assessed in the UK legal system. In null hypothesis testing the null hypothesis is tentatively assumed to be "not guilty". Evidence is then presented concerning the null hypothesis. An evaluation of this evidence leads to a decision of "not guilty" or of "guilty". If there is a "guilty" verdict then the null hypothesis is rejected and because of the mutually exclusive nature of the statistical hypotheses the rejection of the null hypothesis necessarily leads to the acceptance of the alternative hypothesis. On the other hand, finding the null hypothesis "not guilty" does not translate into "innocent". This distinction is particularly important when designing a study, drawing inferences from a study and is particularly important in developing the logic for multiple comparisons. Failure to reject the null hypothesis simply means that sufficient doubt has not been cast on the credibility of null hypothesis and this may simple be because of a small sample size relative to the size of the effect.

V. Statistical Conclusions

The history of statistics is long established. The invention of the decimal system and decimal point in 1585 helped with the uptake of calculating the mean. In 1710 John Arbuthnot looked at the male to female birth ratio and examined whether a 1:1 ratio was tenable. His work, using the Binomial distribution, essentially derived what is now termed a *p*-value [5]. From the late 1800's onwards there was an explosion of statistical theory which gave rise to many of the commonly used statistical tests (e.g. *t*-tests, chi-square, regression, correlation). These discovered tests can be shown to be the best tests possible providing their underpinning assumptions are satisfied. This is very good news. It means that researchers, for the main, do not have to invent new statistical

tests to examine their data. They simply have to carefully design their research and select the most appropriate and best statistical test from the library of statistical tests. The question that arises is, "what do we mean by best?"

A best statistic for any given situation is the one which will have the best or highest chance of correctly arriving at the correct statistical conclusion. If the null hypothesis is false, then the chance of correctly rejecting the null hypothesis is referred to as the power, or the power of the test. If the null hypothesis is true then the power defaults to what is termed, the significance level. The significance level, is denoted by the symbol α.

Imagine a situation where a researcher puts forward a theory encapsulated by the alternative hypothesis, H_1; and this theory is correct. However, suppose that the data collected is not sufficiently convincing to reject the null hypothesis, H_0. That is to say, the true position is that H_0 is false, but there was failure to reject H_0. The occurrence of this is known as an error of the second kind, or equivalently as a Type II error. The probability of a Type II error is denoted by the symbol β. The power of the test is therefore 1 - β.

Now imagine a situation where a researcher puts forward a theory encapsulated by the alternative hypothesis, H_1; but this theory is wrong. However, suppose that the data collected is sufficiently convincing to reject the null hypothesis, H_0, leading to the acceptance of H_1. That is to say, the true position is that H_0 is true, and $H_0$was rejected. The occurrence of this is known as an error of the first kind, or equivalently as a Type I error. The significance level, α, is the largest probability of committing a Type I error that an investigator is prepared to tolerate. Traditionally, the significance level is set at a value of 0.05 (5% significance); this is a default value, but the value is very much context dependent [5].

In general, a Type I error is more damaging to society than a Type II error. For this reason, in carefully designed studies, it is usual to set $\alpha \leq \beta$. So, for instance, a researcher working at α = 0.05, might have power of 80% (β = 0.20) or power of 90% (β = 0.10) or power of 95% (β = 0.05). However, if a researcher is aiming for 99% (β = 0.01) then consideration should be given to reducing α, from its default.

Power is largely dictated by sample size but there are many other considerations.

In summary, a Type I error (or error of the first kind) is said to have occurred if the null hypothesis is rejected when in fact the null hypothesis is true; a Type II Error (or error of the second kind) is said to have occurred if the null hypothesis is not rejected when in fact the null hypothesis is false. The significance level, is the largest probability of committing a Type I error that an investigator is prepared to tolerate, and the power of the test is the probability of rejecting the null hypothesis and depends on the true state of nature.

A third type of error, a Type III error, would occur if the null hypothesis is correctly rejected, the alternative hypothesis correctly accepted but the sample effect (e.g. a negative correlation) is in the opposite direction to the true state of nature (e.g. a positive correlation). In practice this may occur when there are hidden or latent variables which are not accounted for at analysis.

The above narrative has largely avoided making reference to the *p*-value. For a given set of data and a given test statistic the *p*-value is defined to be the largest significance level for which there is failure to reject the null hypothesis. A more detailed exposition of the p-value is given as a separate note [5].

VI. Inductive Inference

It is perfectly fine to draw statistical conclusions based on the analysis of data. It is *hoped* that the result of the statistical test allows an investigator to infer something about the research hypothesis. This is not always possible, as there may be flaws in the design (e.g. poor internal validity, or poor external validity, or poor ecological validity, or poor measurement validity, or extraneous effects [7], and so on).

The argument from the outcome of a statistical test to the probable truth of the scientific hypothesis is an example of inductive *inference*. (Analogous to a proof by induction in mathematics.) We can only re-iterate that it should be noted that if errors occur at any stage of the research then the outcome of the statistical test may have little relevance with respect to the scientific hypothesis.

It is really difficult to do good quality research.

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] Coughlan M. Cronin P, and Ryan F (2007) Step-by-step guide to critiquing research. Part 1: Quantitaive Research, Bristish Journal of Nursing, Vol 16, No 11, 658 – 663.

[2] Santy J and Kneale J (1998) Critiquing quantitative research, Journal of Orthopaedic Nursing, Vol 2, No 2, 77 – 83

[3] Marshall G (2005), Critiquing a research article, Radiography, Vol 11, No 1, 55 – 59.

[4] Astroth KS and Chung SY (2018) Focusing on fundamentals: Reading quantiative research with a crkitical eye, Nephrology Nursing Journal, Vol 32, 32 -38.

[5] 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.

[6] 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