ON THE SUBJECT OF HYPOTHESIS TESTING

ANTONY UGONI B.Sc. (Hons) *
Biostatistician

Abstract: In this paper, the definition of a statistical hypothesis is discussed, and the considerations which need to be addressed when testing a hypothesies. In particular, the p-value, significance level, and power of a test are reviewed. Finally the often quoted confidence interval is given a brief introduction.

Key words: Hypothesis testing, confidence intervals, chiropractic.

1. Introduction

Readers of journal articles are usually aware of what the clinician or experimenter is trying to demonstrate, but sometimes become confused with the resulting barrage of statistics that usually ensue from most experiments. This confusion is magnified due to the fact that the statistician usually tests a hypothesis which is the complete reverse of what the clinician is trying to show. This paper attempts to make the reader more aware of the manner in which the statistician analyses data, and the fundamental tools used to make these analyses.

2. What is a hypothesis?

The Australian Concise Oxford Dictionary defines it as ' ....supposition made as a starting-point for further investigation from known facts....'. In the medical field, hypotheses are proposed and tested constantly, and journals can afford to be choosy about which articles will be published by them. But exactly what are these hypotheses?

This would seem like a trivial issue, if it weren't for the fact that most articles do not state their hypothesis in the 'statistical' sense, and only in an 'optimist' sense. The greatest problem is that hypotheses proposed by the optimist are usually the reverse of the hypothesis proposed by the statistician (1). Take for example a chiropractor who wants to know whether or not smoking increases the risk of back pain. The hypothesis in question here is 'Smoking increases the risk of back pain'. To make sure that the chiropractor covers all possibilities, he/she may choose to hypothesise 'The risk of back pain is altered when smoking', that is, perhaps smoking may increase or decrease the risk of back pain. However, the hypothesis statisticians test is always that no change will occur. That is, smoking does not effect
the incidence of back pain, or, the incidence of back pain is independent of smoking status.

The chiropractor may be interested in showing that risk changes with smoking status, and then interested in the magnitude of that change. Therefore, to show that a change does exist, all the statistician needs to do is show that the occurrence of 'no change' is unlikely. Thus, simply show that the hypothesis 'No change' is untrue, and it follows that the hypothesis 'Some change' is true. It is left to confidence intervals (1) to then give an indication of that change (confidence intervals to be discussed later).

In other words, the clinician sees smoking as a factor which may increase the risk of back pain. Once a difference is observed, it is the statisticians job to determine whether or not that difference is attributed to a chance random variation or a genuine difference. That is, is the observed difference due only to chance?

In general, the statistical hypothesis for (almost) any experiment is the statement that says 'NOTHING HAPPENS'. For this reason, this hypothesis is universally knows as the **null** (1) hypothesis, but shall simply be referred to as the 'hypothesis'. If we can prove that this is untrue, then obviously *something* happens, and it is a matter of estimation to show the size of that change. These steps are outlined below for the general case:

- Collect data to (hopefully) show that change occurs.
- Test the hypothesis 'Change does not occur'.
- Decide whether or not to reject the hypothesis. Rejecting the hypothesis leads to "Change does occur".

3. When do we reject a hypothesis?

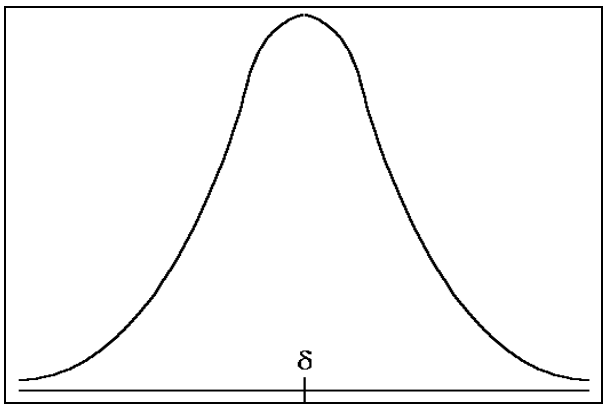
* DEPARTMENT OF SOCIAL AND PREVENTIVE MEDICINE
MONASH UNIVERSITY
COMMERCIAL ROAD, PRAHRAN, VICTORIA, 3181

For most hypotheses, statistical tests are available for use which provide us with a probability known as the p-value (2), and is defined below:

P-value: The probability of observing data as or more extreme than what we observed ('extreme' with respect to the hypothesis).

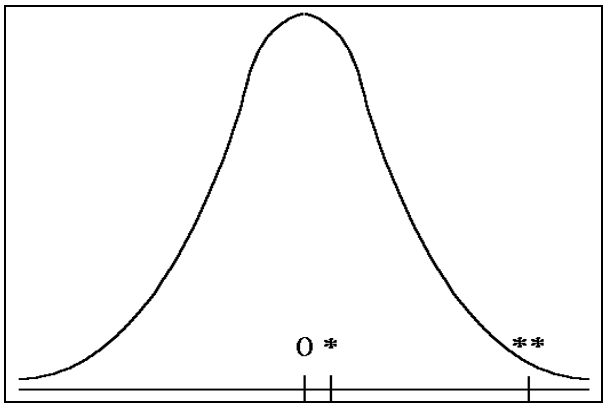
In everyday terms, the probability of our observed result being due simply to random variation.

For example, it may be of interest to study the difference in blood pressure (difference = blood pressure before treatment - blood pressure after treatment) after some treatment (eg. soft tissure massarge) was applied, and to see if the treatment reduced blood pressure. The hypothesis to test is $\delta = 0$ (where δ = mean difference between groups). That is, on average, there is no change between before and after blood pressure values. For the purposes of this example assume that the distribution of the mean difference is 'normal' (1), and is illustrated below.



NB: The 'distribution' of any set of data is a graphical (and mathematical) representation of the way the data is spread, and can be estimated by empirical methods.

Under the hypothesis $\delta = 0$ (the mean difference is zero), the distribution of the mean difference looks like (with some examples of observations * and **) the example below with all the data centred around zero.



Suppose 100 people have had their blood pressures measured before and after treatment, and the mean difference of the sample is calculated and represented by * (for example, a mean difference of 0.9 mm Hg) above. The experimental results suggest little to oppose the hypothesis that the mean difference is zero mmHg. Now suppose the experimental results yield a mean difference represented by ** (for example, a mean difference of 13.2 mm Hg). In relation to the distribution of the data and the hypothesis, this looks like an unlikely event since it lies so far away from the hypothesised mean of zero mmHg. Accordingly, the probability of such an event would be quite small and most people would consider this an aberration. The fact is, however, that the data **was** observed, and the experimenter can only assume that this result observed is representative of the entire population of data. Thus, it is not the observation that looks unlikely, but rather the hypothesis and the distribution it suggests as the true distribution of the entire population.

This argument is the essence of the p-value. In fact, the p-value for ** is the proportion of the total area underneath the curve and to the right of **, and then multiplied by 2 (two sided test, 1). As can be seen, the area to the right of ** is quite small, making the p-value small, and thus the hypothesis seems highly unlikely. On the other hand, the proportion of area to the right of * will be quite substantial, and thus the p-value will be large making it difficult to reject the hypothesis.

In short:

- Conclude no difference exists: observed result is a chance finding.
- Conclude a difference exists: observed result represents a true difference.
  The p-value indicates which is the more likely explanation.

4. How small does the p-value need to be to reject the hypothesis?

The most common criteria used for 'too small p-value' is 0.05. That is, when we observe a p-value smaller than 0.05, then we rejected the hypothesis. After 0.05, 0.1 and 0.01 are also used frequently.

This criteria is known as the significance level (2). Journal articles frequently have statements such as '....the data was significant at the 0.05 level ....'. This means that the data has yielded a p-value small enough to suggest a significant deviation from the hypothesis to be able to reject this hypothesis.
The general notation is

Reject the hypothesis when the p-value $< \alpha$
α = significance level

It must be noted that a possible error can arise when performing any hypothesis test.

Rejecting the hypothesis when the hypothesis is, in fact, true, is known as a 'Type I' error (1). Obviously the statistician would like to minimise the probability of this mistake, or even nullify it altogether. Unfortunately, this mistake will always remain a possibility since the probability of a Type I error occurring is equal to the significance level. That is:

$$P(\text{Type I error occurring}) = \alpha$$

The dilemma is now trying to choose α small enough to make the Type I error unlikely, but large enough to reject the hypothesis if it is false. For example, suppose the 100 blood pressures are measured before and after treatment, and the mean difference is calculated to determine whether or not the population mean difference is equal to zero. Now suppose that the **true** mean difference **is** zero (of course the experimenter will never know this prior to performing the experiment). The statistician chooses $\alpha = 0.05$, and proceeds with the hypothesis test. The choice of $\alpha = 0.05 = 1/20$ can be interpreted in the following way: If the experiment were done 20 times, we would expect to falsely reject the hypothesis once. That is, out of 20 similar experiments, probability dictates that we can expect to give spurious results.

On the other hand, making α very small will mean that we will almost never reject the hypothesis even when it is false.
$\alpha = 0.05$ is commonly accepted for most situations.

5. How can we be sure that we make the correct decision?

There are 2 incorrect decisions that can be made when testing hypotheses. The first, the Type I error has already been discussed above. The second, is useful for the calculation of appropriate sample sizes.

The 'Type II' error (1) is made when the hypothesis is accepted when it is false. The probability of a type II error is denoted as β. The interpretation of $1-\beta$ is given below.

$$P(\text{Type II error}) = P(\text{Accept the hypothesis when it is false}) = \beta$$

$$P(\text{Reject the hypothesis when it is false}) = 1 - P(\text{Accept the hypothesis when it is false}) = 1 - \beta$$

Historically, $1-\beta$ is known as the 'Power' (1) of a test, and although it is not obvious, the equation

$$P(\text{Reject the hypothesis when it is false}) = 1 - \beta$$

is the basis of most sample size calculations (equally important is specifying how large a deviation from the hypothesis we want to detect).

NB: The notation P(event X) is used to denote 'the probability of event X'.

Important to note is that the magnitudes of α and β are usually chosen before the start of the study. That is, the researcher should explicitly state what values are to be used before the study begins. β is used in the design of the study to ensure sufficient numbers of subjects are collected, and α after using the data to choose α.

As a final note, the reader should consider this question. If a significance level of 0.05 is chosen, is there a difference between the p-values 0.051, and 0.049? Implicitly, the answer is yes, and a strict criterion of $\alpha = 0.05$ suggests this also. That is 0.051 suggests we do not reject the hypothesis, and 0.049 suggests we do reject the hypothesis. However, the difference between 0.051 (51 per 1000), and 0.049 (49 per 1000) is 0.002 and the change from rejection to acceptance of the hypothesis could easily be due to a minor chance fluctuation in the data. Thus, p-value's in this region should be considered with care, and, although interpreted as small, should always be regarded in terms of the implications of rejecting or not rejecting the particular hypothesis.

6. What is a 'Confidence Interval'?

Hypothesis testing provides an indication of the likelihood of a hypothesis. This area of statistics is referred to as inferential (2). Another area of statistics is estimation (3). A sample mean difference is an estimate of the population mean difference. In the above notation, the sample mean difference δ, is an estimate of the population mean difference Δ.

More and more journals are demanding their writers to produce confidence intervals (1). This is due to the fact that they are easier to interpret, and much more appealing to the eye of a non-statistician (and certainly a large number of statisticians). They represent a range which the statistician is 'confident' the true value of the statistic lies within.

For example, a chiropractor may be interested in the mean change in heart rate (beats/minute) δ after C1

manipulation for male patients aged between 20-30 years of age, and wants to test the hypothesis that δ = 5 beats/minute. Suppose the hypothesis is successfully rejected due to a small p-value. The chiropractor is now interested in the best estimate of the mean change in heart rate.

The reader should now be made aware of variation between sampling. For example, to estimate Δ, the chiropractor may take a random sample of 20 such men and calculate the sample mean to be (say) 11. Another random sample may yield an estimate of 13, and another may give 6, etc. Obviously the random sample will produce a sample mean change which is not exactly equal to the population mean change.

A simple experiment to illustrate this is to try and estimate the probability of heads turning up on the flip of a coin. Flip the coin 10 times, and count the number of heads. The estimated probability of heads is then the number of times heads appear divided by 10. Ten trials of this experiment gave the following number of heads: {2,5,6,5,4,4,4,8,5,4}, where we expected 5 heads. Had we not known the true probability (1/2), it would be difficult at best to try and be confident about quoting a single number for the probability of heads.
To overcome this problem, statisticians have devised the 'Confidence Interval'. Instead of quoting one number for the estimate of δ,statisticians quote an interval (or range) of numbers which they are confident that δ lies within.

How confident are the statisticians?

Within a number of assumptions (most of them well founded), the statistician can quote any degree of confidence desired, but the most common degree of this confidence used is 95%. For example, the 95% confidence interval for δ may have been 6 to 13 beats per minute. That is, we are 95% confident that Δ (the true population mean) lies between 6 and 13 beats per minute. The stricter interpretation of a (say 95%) confidence interval is: for 100 independently calculated 95% confidence interval's, we can expect 95 of them to encompass the true value of Δ, and 5 of them to not encompass Δ within their bounds. A simulation of 1000 such confidence intervals, where the known mean is zero, had 956 (95.6%) intervals where zero was encompassed.

7. Discussion

Without delving into the mathematical detail, the foundations of hypothesis testing and a basic introduction to confidence intervals were discussed. While most readers of journal articles are unaware of the design of the experiment to begin with, they should always keep in mind the question that asks: Is the design competent, and the sample size large enough to test the hypothesis proposed? Large sample sizes lead to large powers, whereas small sample sizes are rarely ever able to reject the hypothesis in question. A test to determine which of treatment A and treatment B is more beneficial (for example) will never discriminate between the two if a small sample size is used, thus a possibly new and better method may be left unused.

The interpretation of a p-value was mentioned, and basically the rule is, 'The smaller the p-value, the less likely the observed result was a chance finding.

The reader would do well to remember that the conviction of a criminal is analogous to hypothesis testing. The alleged criminal is innocent until proven guilty. In other words, we assume the alleged criminal has done nothing (no change), and then use all the evidence to show that innocence is highly unlikely (reject hypothesis).

Acknowledgment

The author wishes to express gratitude to Andrews Forbes and Belinda Gourlay for comments and recommendations on every draft, and John Drinkwater for his chiropractic insight.

References

1. Brown, B. and Hollander, M. Statistics: A Biomedical Introduction.
   Wiley, New York, 1977.
2. Armitage, P. and Berry, G. Statistical Methods in Medical Research.
   Blackwell Scientific Publications, Oxford, 1988
3. Goldman, R. and Weinberg, J. Statistics: An Introduction.
   Prentice Hall, New Jersey, 1985.