Statistical Analysis of Survey Data by Using Hypothesis Testing

Myo Khaing, Nan Sai Mon Khan
University of Computer Studies, Yangon
myokhaing.ucsy@gmail.com, myo_khaing@yahoo.com

Abstract

A common goal for a statistical research project is to investigate causality, and in particular to draw a conclusion on the effect of changes in the values of predictors or independent variables on dependent variables or response, there are two major types of causal statistical studies; experimental studies and observational studies. In both types of studies, the effect of differences of an independent variable (or variables) in the behavior of the dependent variable are observed. The term "chi-square" refers both to a statistical distribution and to a hypothesis testing procedure that produces a statistic that is approximately distributed as the chi-square distribution. Whether analyzing null-hypothesis is or not by using chi-square entirely depends on the significant level (alpha) and sample size. Whenever we make a decision based on a hypothesis test, we can never know whether or decision is correct. There are two kinds of mistakes we can make: (1) we can fail to accept the null hypothesis when it is indeed true (Type I error), or (2) we can accept the null hypothesis when it is indeed false (Type II error). This paper tries to reduce the chance of making either of these errors by adjusting between the significant level (alpha) and the minimum sample size needed.

1. Introduction

A test is a statistical procedure to obtain a statement on the truth or falsity of a proposition, on the basis of empirical evidence. This is done within the context of a model, in which the fallibility or variability of this empirical evidence is represented by probability. In this model, the evidence is summarized in observed data, which is assumed to be the outcome of a stochastic, i.e., probabilistic, process; the tested proposition is represented as a property of the probability distribution of the observed data.

1.1 Related Works

The first published statistical test was by John Arbuthnot in 1710, who wondered about the fact that in human births, the fraction of boys born year after year appears to be slightly larger than the fraction of girls [6].

One of the first statistical procedures that come close to a test in the modern sense was proposed by Karl Pearson in 1900.Thiswas the famous chi-squared test for comparing an observed frequency distribution to a theoretically assumed distribution. This distribution can therefore be used to calculate the probability that, if the hypothesis holds, the test statistic will assume a value equal to or larger than the value actually observed.

The idea of testing was further codified and elaborated in the first decades of the twentieth century, mainly by R. A. Fisher [7]. In his significance tests the data are regarded as the outcome of a random variable X (usually a vector or matrix), which has a probability distribution which is a member of some family of distributions; the tested hypothesis, also called the null hypothesis and the significance of the given outcome of the test statistic is calculated as the probability. The significance probability is now often called the p-value (the letter p referring to probability). With Fisher originates the convention to consider a statistical testing result as `significant' if the significance probability is 0.05 or less [7]. A competing approach was proposed in 1928 by Neyman and Egon Pearson [8].

Neyman and Egon Pearson (the son of Karl) [8] criticized the arbitrariness in Fisher's choice of the test statistic and asserted that for a rational choice of test statistic one needs not only a null hypothesis but also an alternative hypothesis. They formalized the testing problem as a two decision problem. Denoting the null hypothesis by H0 and the alternative H1, the two decisions were represented as `reject H0' and `do not reject H0' Two errors are possible: rejecting a true H0, and failing to reject a false H0. Neyman and Pearson conceived of the null hypothesis as a standard situation, the burden of proof residing with the researcher to demonstrate (if possible) the untenability of this proposition. Correspondingly, they called the error of rejecting a true H0 an error of the first kind and the error of failing to reject a false H0 an error of the second kind. Errors of the first kind are considered more serious than errors of the second kind. The probability of correctly rejecting H0 If H1 is true, which is 1 minus the probability of an error of the second kind, given that the alternative hypothesis is true, they called the power of the test. Neyman and Pearson proposed the requirement that the probability of an error of the first kind, given that the null hypothesis is indeed true, do not exceed some threshold value called the significance level usually denoted by 'alpha'. Further they proposed to determine the test so that, under this essential condition, the power will be maximal.

In the Neyman-Pearson formulation, we obtain richer results at the cost of a more demanding model. In addition to Fisher's null hypothesis, we also need to specify an alternative hypothesis; and we must conceive the testing problem as a two-decision situation. This led to vehement debate between Fisher on the one hand, and Neyman and E. Pearson on the other. This debate and the different philosophical positions are summarized by Hacking [6] and Gigerenzer et al. [9], who also give a further historical account.

Examples of this hybrid character are that, in accordance with the Neyman-Pearson approach, the theory is explained by making references to both the null and the alternative hypotheses, and to errors of the first and second kind (although power tends to be treated in a limited and often merely theoretical way), whereas in the spirit of Fisher statistical tests are regarded as procedures to give evidence about the particular hypothesis tested and not merely as rules of behavior that will in the long run have certain (perhaps optimal) error rates when applied to large numbers of hypotheses and data sets. Lehmann [10] argues that indeed a unified formulation is possible, combining the best features of both approaches.

Instead of implementing the hypothesis test as a `reject/don't reject' decision with a predetermined significance level, another approach often is followed: to report the p-value or significance probability, defined as the smallest value of the significance level at which the observed outcome would lead to rejection of the null hypothesis. Equivalently, this can be as the probability, calculated under the null hypothesis, of observing a result deviating from the null hypothesis at least as much as actually observed result. This deviation is measured by the test statistic, and the p-value is just the tail probability of the test statistic. For a given significance level a, the null hypothesis is rejected if and only if p<=a [1].

2. Hypothesis Testing

The theory of hypothesis testing is concerned with the problem of determining whether or not statistical hypothesis, that is, a statement about the probability distribution of the data, is consistent with the available sample evidence. The particular hypothesis to be tested is called the null hypothesis and is denoted by H0. The ultimate goal is to accept or reject H0.

In addition to the null hypothesis H0, one may also be interested in a particular set of deviations from H0, called the alternative hypothesis and denoted by H1. Usually, the null and the alternative hypotheses are not on an equal footing: H0 is clearly specified and of intrinsic interest, whereas H1 serves only to indicate what types of departure from H0 are of interest.

2.1 Types of Hypothesis

The art of statistics is in finding good ways of formulating criteria, based on the value of one more statistics, to either accept or reject the null hypothesis H0. It should be noted that H0 and HA can be almost anything, and as complicated or as simple as we wish. If a hypothesis is stated such that it specifies the entire distribution, we call it a simple hypothesis. Otherwise, we call it a composite hypothesis. As you might imagine, more rigorous tests can be done with simple hypotheses, because they specify the entire distribution, from which probability values can be computed.

There are two hypotheses that are possible:

Null Hypothesis: The statement being stated in a test of significance is called the null hypothesis. The test of significance is designed to assess the strength of the evidence against the null hypothesis. Usually the null hypothesis is a statement of “no effect” or “no difference”. The null hypothesis is usually denoted by H0.

Alternative Hypothesis: The statement that we suspect to be true. Or the statement that we wish to conclude. This alternative hypothesis is usually denoted by Ha or H1.

Steps to do Hypothesis Testing:

1. Formulate the Null hypothesis and Alternative hypothesis.
2. Specify the level of significance (Commonly used: = 0:05 or 0.01).
3. Determine the appropriate test statistic to use.
4. Define your rejection rule (Not needed if you decide to use the p-value).
5. Compute the observed value of the test statistic (Compute the p-value).
6. Write your conclusion. [3]

2.2 Type I and Type II errors

In any testing situation, two kinds of error could occur:

Type I (false positive): We reject the null hypothesis when it’s actually true.

Type II (false negative): We accept the null hypothesis when it’s actually false.

The probability of committing a Type I error is typically denoted α, and the probability of a Type II error is denoted β.

α: the probability of making a Type I error (false positive).

β: the probability of making a Type II error (false negative).

"α" is often called a significance level or sensitivity. Typically, we try to fix an accepted level, α of Type I error, and go on to find ways of minimizing the level of Type II error, β.

The statistical power of a test is defined as (1 − β). We usually want to maximize the power of our test in order to detect as many significant signals from our data as we possibly can [2].

The Probability of the Null Hypothesis

The first misinterpretation is to view a p-value as the probability that the results occurred because of sampling error or chance fluctuations. For example, p=0.05 is interpreted to mean that there is a probability of only .05 that the results were caused by chance. However, this interpretation is completely erroneous because (1) the p-value was calculated by assuming that the probability is 1.0 that any differences were the result of chance and 92) the p-value is used to decide whether to accept or reject the idea that the probability is 1.0 that chance caused the mean difference. A p-value of .05 means that, if the null hypothesis is true, the odds are 1 in 20 of getting a mean difference this large or larger and the odds are 19 in 20 of getting a smaller mean difference. *However, there is no way in classical statistical significance testing to determine whether the null hypothesis is true or the probability that it is true.*

2.3 The Probability of Results Being Replicated

A second misinterpretation is that the p-value represents the confidence a researcher can have that a given result is reliable or can be replicated. Basically, this argument is that the complement of the p-value yields the probability that a result is replicable or reliable, eg 1-.05=.95 probability that results can be replicated. This misinterpretation probably comes from a notion that a statistically significant difference in sample means suggests that the samples are from different hypothetical populations and future samples drawn from these different hypothetical populations will therefore yield q=equivalent results. *However, nothing in classical statistical significance testing says anything about the probability that the same results will occur in future studies.* Replication results is a function of how exactly the method is repeated, and some aspects, such as the time of measurement, clearly cannot be identical to those of the original study.

2.4 The Probability of Results Being Valid

The third and most serious misinterpretation of classical statistical significance testing is that it directly assesses the probability that the research (alternative) hypothesis is true. For example, a p-value of .05 is interpreted to mean that its complement, .95, is the probability that the research hypothesis is true. Related to this misinterpretation is the practice of interpreting p-values as a measure of the degree of validity of research results, i.e., a p-value such as p<.0001 is "highly statistically significant" or "highly significant" and therefore much more valid than a p-value of, say, .05. Again, such a practice is inappropriate. Although it is true, for example, that the greater the difference between group means the greater the chance of obtaining a small p-value, and it is true that such a result may be rarer given the null hypothesis a statistically significant result cannot properly be construed as a more valid result for at least two reasons.

First, a statistical test is not a complete test of a research hypothesis. Instead it examines only one of many possible operations of a research hypothesis. Thus, it is improper to infer that the research hypothesis is valid without testing and support from a representative sample of operations. Second, a variety of threats to drawing valid inferences are not addressed by statistical tests (Cook and Campbell 1979). *In any event, rejection of the null hypothesis at a predetermined p-level supports the inference that sampling error is an unlikely explanation of results but gives no direct evidence that the alternative hypothesis is valid.*

2.5 Sample Size and Probability of the Research Hypothesis

Moreover, because effect size is a measure of the strength of the relationship and large effects are more likely to be replicated than small ones, researchers should have more confidence in the study with the smaller sample.

3. Parametric versus Non-parametric Tests

In general, there are two kinds of statistical tests. Classical statistics mostly deals with parametric tests.

These are tests which assume some sort of model for the underlying distribution. Many of the statistical distributions used in these tests assume that the data is drawn from a normal distribution. Given this assumption, much can be derived about the distribution of the observations themselves.

Non-parametric tests do not assume any kind of underlying probability distribution. This can be very useful in cases where it would be very hard to justify that the data are normally-distributed (or if we know it's just plain not true). Many non-parametric tests can quite powerful simply by considering the rank order of the observations [2].