Misuse of P-Value in the Health Sciences

Ram C. Dahiya
Department of Mathematics and Statistics
Old Dominion University Norfolk, VA 23529, U.S.A.

Abstract Statistical testing in the health sciences has given too much importance to p-value while diminishing the importance of power of the test. Any statistical test has two types of errors, Type I with probability denoted by α and Type II with probability denoted by β. Attention should be given to both of these errors while comparing different tests or when using any specific test. A p-value of less than .05 has become almost essential for rejecting a null hypothesis H_0 in favor of some alternative hypothesis H_1. Because of the desire for a small p-value, without consideration to the power of the test, a recent paper by Sadek et al. (1999) reaches a wrong decision in rejecting the survival benefit of a potentially useful treatment for Ebola hemorrhagic fever. Here we demonstrate the drawbacks of this type of practice in statistical testing, especially for small samples.

1. INTRODUCTION

For any statistical test testing the null hypothesis, H_0, against an alternative hypothesis, H_1, two types of errors are given by

$$\alpha = P(Reject\ H_0 | H_0)\ ,$$

and

$$\beta = P(Reject\ H_1 | H_1)$$

where α is known as Type I error or the level of significance and β is known as Type II error. The power of the test is given by $(1-\beta)$. For a given test, not depending on α and β, both types of errors decrease as the sample size n increases. In practice, researchers usually fix α to specify a test and ignore looking at the power of the test. In such a situation, β decreases as n increases but α remains fixed. This is not good practice and one should look at both α and β for a given n before fixing α. In the case of a composite alternative H_1, one can look at a few reasonable values of β before fixing α.

In health sciences, researchers usually simply report the p-value of the test. P-value in general is defined to be the smallest level α for which the null hypothesis can be rejected for data at hand. There are, of course, different interpretations of p-value (cf. Gibbons and Pratt (1975)). Birnbaum (1962) defines p-value as the "observed level of significance". Prevalence of misunderstanding about the meaning of p-vlaue is high among health researchers. Freeman (1993) reports the results of a multiple choice test about the meaning of p-value given to 397 doctors, dentists, and medical students. Only 19% of the respondents chose the correct answer out of four possible answers. Note that, if respondents were to make their choice at random even then we would expect 25% of them marking the correct answer. Freeman (1993) states that, "the current widespread practice of using p-values as the main means of assessing and reporting the results of clinical trials cannot be defended".

Practice by some health researchers to give p-values alone without giving any data or the details about the test used is prevalent and must be discouraged by the research journals in health-related areas. Journal editors should require authors to provide power for the given p-values; for at least a few specific alternatives. The alternative of providing confidence intervals does not solve the problem, as this leads to using a fixed α, usually .05, for any n.

Computations of p-value for most statistical tests is available in any statistical software but one needs the help of a well-trained statistician or biostatistician to look at the power of the test being used. Furthermore, altrnatives, in general, are composite and hence one needs to look at the behaviour of power for a few specific alternatives. For the case of comparing two populations, the alternatives can easily be defined but for more than two populations even specifying the alternatives is a difficult problem. This has lead the researchers to the risky practice of depending on p-value alone.

An example of misuse of p-values is when researchers in health sciences compare p-values of an asymptotic test with an exact test for small samples and then declare the asymptotic test to be superior because it gave a lower p-value, without any regard to power

comparisons of the two tests. Asymptotic tests are used for a small sample size without giving any consideration to the underlying assumptions. Hence, the conclusions based on these types of tests may be invalid even if one gets a small p-value. Exact tests in some situations provide a much better alternatives to the asymptotic tests. Here, we are going to demonstrate the drawbacks of this type of practice in statistical tests, especially for small samples. To give another, more extreme example, an adhoc test which always accepts H_0 has a p-value of zero but it also has zero power giving Type II error $\beta=1$. We are going to consider situations where using a value of α as high as 0.3 is justified because of power considerations. However, it will be hard to find a case in the literature where H_0 is rejected when the p-value is as high as 0.3. Here, we are going to show that in a recent study (Sadek et al. (1999)) in the Journal of Infectious Diseases, wrong conclusions were drawn regarding the effectiveness of a treatment for Ebola hemorrhagic fever because the authors did not get a desired low p-value for a small sample size.

The proper way of carrying out any test for a given n is to look at both α and β and then decide about the level of the test. The level of the test may also depend on the situation at hand. In the case of testing the effectiveness of the only treatment available (e.g. treatment for Ebola in Sadek et al. (1999)) for a disease, one should be willing to accept a much higher value of α as compared to the test comparing the effectiveness of two different treatments. Furthermore, in the case of a discrete null distribution of the test statistic, the commonly used level of 0.05 may not even be attainable. Because β decreases as n increases and tends to zero as $n\to\infty$ for a fixed alternative, it is only natural to fix a lower value of α for higher sample size and to accept much higher values of α than the nominal 0.05 for smaller samples.

Fisher (1935) gives a Tea Tasting example where a woman claiming to distinguish if milk or tea was added to the cup first was given 8 cups with four of each kind. She correctly guessed three of the four cups having milk added first. Using the hypergeometric distribution for the number of correct guesses under the null hypothesis of random selection, the p-value is given by 0.243. The lowest possible value for p-value, when all four are correctly identified, is 0.014. If one were to use a level of 0.01, the null hypothesis will never be rejected even if all the guesses are correct. This example underlines the problem with desiring a small level of significance in small samples.

2. TESTING FOR BINOMINAL PROBABILITY

Let us consider a rare disease and assume that only 1% of the general population has this disease. We are interested in testing if a specific group, defined by specific diet or some other characteristic, has a lower prevalence than the general population. In a sample of size 250 from the specific group, let X be the number of people with the disease. Then, X has binomial (n,θ) distribution with $n=250$ and $\theta=$ probability of a person having the disease. We are interested in testing the following hypothesis:

$$H_0:\theta=.01 \quad , \quad H_1:\theta=.005.$$

Note that the alternative states that the prevalence rate in the treatment group is 50% of the general population. We are going to use an exact binomial test based on X and will obviously reject H_0 for small values of X. Let us consider the power and two types of errors for the following two tests given in Table 1.

Test 1: Reject H_0 if $X=0$.
Test 2: Reject H_0 if $X\le 1$.

Table 1

	α	β	Power
Test 1	.081	.714	.286
Test 2	.286	.356	.644

It is obvious from Table I that the smallest p-value for this case is .081. In fact, you need a sample of size 300 to get the smallest possible p-value down to .05. However, looking at the power and Type II error, Test 2 with $\alpha=0.286$ is more reasonable and there is nothing sacred about using $\alpha=.05$. Here, we are talking about an exact test for detecting 50% improvement in the prevalence rate of the disease. Even for $n=1000$, assuming $(X-n\theta)/\sqrt{n\theta(1-\theta)}$ converging to standard normal, a test of size $\alpha=0.1$ has power $=0.666$ and $\beta=0.334$, indicating that one should use a test of size $\alpha>0.1$.

Freeman (1993) states, in conclusion, that one always should require a p-value of 0.001 or less in large samples ($n>200$, say) before declaring anything significant. The example above demonstrates the fallacy of this kind of rule for fixing a significance level solely based upon the sample size. The

significance level should be fixed by looking at both types of errors and any other adhoc rule is bound to have problems under certain circumstances.

3. TREATMENT FOR EBOLA HEMORRHAGIC FEVER

Sadek et al. (1999) provides statistical analysis of an epidemic of Ebola hemorrhage fever in 1995 in the Democratic Republic of the Congo. The age of a patient was found to be a significant factor affecting the probability of survival. Of 310 cases, 250 (80.8%) patients died. A treatment involving whole blood transfusion (BT) from convalescent patients was attempted on 8 patients. Although 7 of 8 treated patients survived, Sadek et al. (1999) concluded the following about BT treatment: "No statistical evidence of a survival benefit of transfusion of blood from convalescent patients was evident after adjusting for
age, sex, and days since onset of symptoms (p-value=0.1713)". Here, we show that there is indeed statistical evidence that BT treatment increases the chance of survival. Part of the problem with the conclusions in loc cit. is that they are looking for a small p-value without giving any consideration to the power of the test.

The data regarding the BT patients and conditional probability of survival (conditioned on number of days since onset of BT treatment) based on the control group of 32 patients (matched for age, sex, and days since onset of symptoms) are given below in Table 2 (derived from Tables 1 and 2 of Sadek et al. (1999)).

Table 2

Data for Blood Transfusion Patients of Ebola
(From Tables 1 and 2 of Sadek et al. (1999))

Patient Number (i)	Days Onset to BT	Age (yrs)	Status	Conditional Probability of Survival from Control Group (p_i)
1	4	48	D	0.250
2	7	27	A	0.412
3	9	54	A	0.636
4	11	12	A	0.875
5	11	40	A	0.875
6	13	15	A	1
7	13	25	A	1
8	15	44	A	1

A: Alive D: Dead

The crucial question is to test the following null hypothesis:

H_0: BT treatment has no affect on survival probability

H_1: BT treatment improves the probability of survival, p_i, by $m\%$, i=1,2,..,8. ($p_{i1}=(1+m/100)p_i$)

Because patients are given BT treatment at different times from onset of the symptoms and because the conditional probability of survival, given a patient survives t days, depends on t, even under H_0 the probability of survival for different patients is different.

The use of t-test to test if the observed conditional probabilities can be a random sample from a distribution with mean 0.875 (average of observed survival probability of 8 patients) in Sadek et al. (1999) is erroneous because the patients cannot be thought of as a random sample with the same probability of survival. By their own admission, the Bernoulli observations regarding success or failure of BT are not identically distributed. There is no basis for using a t-test in this situation. Here, we propose to use the exact test given below.

Let

$$X_i=\begin{cases}1 & \text{if ith patient survives}\\ 0 & \text{otherwise.}\end{cases}$$

Now X_i is Bernoulli$(1,p_i)$ under H_0, where p_i is the conditional probability of survival given in Table 2. If we define

X = total number of survivals of 8 patients,

then note that the distribution of X is not binomial because $X_1,X_2,..,X_8$ are not identical Bernoulli random variables.

The joint probability density function (p.d.f.) is given by

$$f(x_1,x_2,..x_8)=\prod_{i=1}^{8} p_i^{x_i}\, q_i^{1-x_i} \quad , \quad q_i=1-p_i, \quad i=1,2,..8.$$

obviously, an exact test based on X will reject H_0 for large values of X. We consider the following two tests:

Test 1: Reject H_0 if $X=8$
Test 2: Reject H_0 if $X\geq 7$.

For Test 1, we have

$$\alpha_1=P(X=8|H_0)=\prod_{i=1}^{8} p_i=0.0502,$$

on using p_i's from Table 2.

Similarly, for Test 2, we have

$$\alpha_2=P(X\geq 7|H_0)=\alpha_1+\alpha_1\sum_{1}^{8} (q_i/p_i)=0.3244.$$

For computations of power, we need to fix different values of m and we find that $m=40$ gives two types of errors for Test 2 to be close to each other. These comparisons are given in Table 3.

Table 3

Comparisons of level and power ($p_{it}=1.4p_i$)

Test (i)	α_i	β_i	Power
1	.0502	.82	.18
2	.3244	.33	.67

First note that no treatment can do better than all 8 patients surviving. Test 1 accepts the alternative that the treatment improves the probability of survival only if all 8 patients survive and even then its level $\alpha=.0502$ is slightly over the minimum p-value desired by most researchers in health sciences. On the other hand, looking at the power, we should use Test 2 instead of Test 1 which gives approximately same value for two types of errors for an alternative with 40% improvement in survival probability of a treated patient. Note that for H_1 with less than 40% improvement in survival probability of a treated patient, the type II error will be higher than 0.33.

On using Test 2 and $\alpha=0.33$, we reject H_0 (p-value = 0.3244) and conclude that there is merit in the BT treatment and it should not be ignored. If an Ebola epidemic is to occur again, the BT treatment should be tried and tested again, perhaps on a larger number of patients.

For small sample sizes, this example suggests that there is nothing wrong in rejecting H_0 even if a p-value is as high as 0.32 and one should not always look for p-values lower than .05. Furthermore, attention to Type II error is more important than Type I error in this case and there is no reason to expect $\alpha<\beta$. Making aType II error in this example would mean rejecting an effective treatment for Ebola while making a Type I error results in accepting a non-effective treatment. Obviously, in such situations we should not look for small p-values, which may result in very high Type II error.

In a life-threatening situation such as an Ebola epidemic, it is irresponsible to reject a potentially life saving treatment like BT simply because the statistical test does not achieve a small p-value. Afterall, only 12.5% of the treated patients died while 80.8% of all the cases resulted in death.

4. REFERENCES

Birnbaum, A., On the foundations of statistical inference, Journal of the American Statistical Association, 57, 269-306, 1962.

Fisher, R.A., The Design of Experiments. Edinburgh: Oliver and Boyd, 1935.

Freeman, P.R., The role of p-values in analysing trial results, Statistics in Medicine, 12, 1443-1452, 1993.

Gibbons, J.D. and Pratt, J.W., P-values interpretation and methodology, The American Statistician, 29, 20-25, 1975.

Sadek, F.S., Khan, A.S., Stevens, G., Peters, C.J., and Ksiazek, T.G., Ebola hemorrhagic fever, Democratic Republic of the Congo, 1995: Determinants of Survival, The Journal of Infectious Diseases, 179, 24-27, 1999.