Shane Allua, PhD, and Cheryl Bagley Thompson, PhD, RN

Hypothesis Testing

This article is the 14th in a multipart series designed to improve the knowledge base of readers, particularly novices, in the area of clinical research. A better understanding of these principles should help in reading and understanding the application of published studies. It should also help those involved in beginning their own research projects.

Hypothesis testing is the method for determining the probability of an observed event that occurs only by chance. If chance were not the cause of an event, then something else must have been the cause, such as the treatment having had an effect on the observed event (the outcome) that was measured. This process of testing a hypothesis is at the heart of most statistical analyses for individual research projects.

In experimental research studies (see part 3 in this series[1]), the investigator separates the subjects into two groups and does something to one group, such as administer a treatment or drug, that is not done to the other group. The results for the treatment and control groups are compared. If the results are different, the investigator seeks to conclude, with some degree of certainty, that the difference is large enough to attribute it to a systematic difference in groups (ie, treatment effect) as opposed to something other than the treatment (ie, random sampling error).

Although in most cases the investigator is trying to find a difference between the two groups, by convention, the null hypothesis is the hypothesis that is tested. The null hypothesis states that there is no difference between the groups being compared. Given a control and a treatment group, the null hypothesis states that the treatment had no effect on the dependent variable (the outcome, eg, mortality rate). Thus, on average, the mortality rate is the same for both control and treatment groups.

Hypotheses

In hypothesis testing, the research question of interest is simplified into one of two possible hypothesis types: the null hypothesis, denoted H_0, or the research hypothesis (sometimes called the alternative hypothesis), denoted H_1. In most cases, it is easier to reject a statement that is false (null hypothesis); therefore, we generally assume the opposite of our research question of interest. We then test to determine whether we can reject the null hypothesis to provide support for our research hypothesis. Hypothesis testing does not prove that the research hypothesis is true but rather suggests that the research hypothesis is plausible.

Steps for Hypothesis Testing

In general, regardless of the research question, there are four hypothesis testing steps:

1. Formulate the null and research hypotheses.
2. Select a significance level.
3. Calculate the test statistic.
4. Analyze and state conclusion based on H_0 (null hypothesis).

Step 1: Formulate the Null and Research Hypotheses

Although we initially assume the null hypothesis to be true, the investigator seeks to answer his or her research question (see part 1 in this series[2]). For example, in a clinical trial of a new drug, our research question might ask, "Is the new drug more effective (ie, decrease mortality) than a placebo (sugar pill)?" We would define our null and research hypotheses as

- H_0: The average mortality rate does not differ between the treatment and control groups
- H_1: The average mortality rate does differ between the treatment and control groups

Research hypotheses refer to relationships suspected in the entire population of interest. In this example, the researcher would hypothesize that the new drug is more effective in all transported patients meeting the inclusion criteria (population of interest). However, when an investigator actually tests a hypothesis, they are running the statistical analysis in only a small sample and then inferring these findings to the population of interest.

Step 2: Select a Significance Level

The significance or alpha level (α) establishes the probability that the investigator is willing to accept that he or she has incorrectly rejected the null hypothesis. In other words, given $\alpha = .05$, the investigator is willing to accept that his or her decision to reject the null hypothesis will be wrong 5% of the time (ie, say that the treatment significantly decreases mortality when actually it does not). The most common alpha levels are .05 and .01.

The significance level is also the criterion level used to define the cutoff probability (of seeing a sample mean, for example) that is so extreme that we consider it very unlikely that we would see a sample mean associated with that probability by chance. This cutoff probability or critical value is used to determine whether the null hypothesis should be rejected and is discussed in detail in step 4. In essence, the alpha level is used to establish statistical significance of the findings and should thus be established a priori (before you actually run the statistical test).

Step 3: Calculate the Test Statistic

Having established the significance level to test the hypothesis, the investigator can now collect data and conduct the statistical test that has been selected. The process for calculating a test statistic is beyond the scope of this paper. The reader is advised to consult with a statistician for further detail and may wish to consult other statistical references[3] on the topic.

Step 4: Analyze and State Conclusion based on H_0

The test statistic calculated in step 3 is then compared with the critical value established, in part, by the significance level selected in step 2. The critical value defines the set of values of the test statistic (the critical region) for which the null hypothesis should be rejected. In other words, values of the obtained test statistic that equal or exceed the critical value and therefore fall within the critical region are considered unlikely to occur as a result of chance or sampling error alone. The investigator can then conclude, with some confidence, that there is a "statistically significant difference" or that the extreme value of the test statistic was attributable to the treatment.

We use the probability value (*p*-value) of a hypothesis test to determine the probability of obtaining an equal or more extreme value of the test statistic than that calculated in step 3 because of chance alone, given a true null hypothesis. We compare the significance level defined in step 2 ($\alpha = .05$) with the *p*-value obtained with our test statistic in step 3. If the *p*-value obtained is smaller than .05 ($p < .05$), we reject H_0. If the *p*-value is larger than .05 ($p > .05$), we fail to reject H_0. Thus, the probability that a sample drawn from a specified population (ALL transported patients meeting your inclusion criteria) will produce a test statistic value greater than or equal to the critical value is less than 5%. Only 5% of samples drawn from a population produce test statistic values as extreme or more extreme than your obtained test statistic. Smaller *p*-values provide evidence that the null hypothesis is unlikely to be true, providing support for the research hypothesis.

An example of hypothesis testing would be evaluating the effect of a drug for controlling hypertension. The investigator could examine blood pressure as the outcome variable. The blood pressure of an individual is related to a variety of factors, so it will not be constant over time. Blood pressure for a given individual may vary naturally (or only by chance) 10 or more mmHg from time 1 to time 2. So if blood pressure has a 10-mmHg drop in blood pressure at time 2, the investigator needs to ask whether this is attributable to normal variation or the to the effect of the drug. We can use the four hypothesis testing steps to assist the investigator in answering this question.

Step 1: Formulate the Null and Research Hypotheses

H_0: The average change in blood pressure does not differ from time 1 to time 2

H_1: The average change in blood pressure does differ from time 1 to time 2

Step 2: Select a Significance Level

After careful consideration of the consequences of committing a type I and type II error, we determined that setting a significance level of $\alpha = .05$ is appropriate for our research project. This means that we are willing to accept that, 5% of the time, we will incorrectly reject the null hypothesis and conclude that the investigational drug significantly decreased blood pressure when it did not.

Step 3: Calculate the Test Statistic

Given that we are comparing the average change in blood pressure from time 1 to time 2 for our entire sample, the appropriate test statistic for our research design is a repeated measures or dependent samples *t*-test.[3] Based on data analysis, our obtained test statistic value is $t_{obt} = .90$.

Step 4: Analyze and State Conclusion Based on H_0

We can now compare our obtained test statistic value, $t_{obt} = .90$, to the critical value of *t*, $t_{crit} = 1.697$, determined, in part, by the significance level set in step 2. Because $t_{obt} < t_{crit}$ ($.90 < 1.697$) and therefore does not fall within the critical region, we cannot reject the null hypothesis. The investigator would conclude that there is no significant difference in average change in blood pressure from time 1 to time 2. Any fluctuations in average change in blood pressure differ by no more than would be expected by chance or on the basis of sampling error.

Potential Errors

When doing hypothesis testing, you can make one of two types of errors. You can conclude that there is a difference between the two groups, when there actually is no difference, or you can conclude that there is no difference between the two groups when actually there is a difference (fail to measure a difference). The first error is called a type I (or alpha) error, and the second is called a type II (or beta) error.

Type I and II errors are important to the investigator because he or she can make decisions in the study design and data analysis that will decrease one or the other of these errors. This is particularly important because the investigator cannot know whether they have made either of these errors when interpreting the data.

Table 1.

		Truth	
		H_0 No Treatment Effect	**H_1 Treatment Effect**
Decision	H_0 No Treatment Effect	Correct	Type II error
	H_1 Treatment Effect	Type I error	Correct

An additional difficulty is that, as the chance of a type I error decreases, the chance of a type II error increases and vice versa. Given our decision to either reject or not reject the null hypothesis, there are two alternatives for reality or truth. The following table gives the four possible results for any hypothesis test.

Type I Errors

Type I errors reflect overconfidence in a finding of differences between two groups. The frequency of type I errors is directly related to the alpha set by the investigator (step 2). For example, if the investigator sets his or her alpha at .05, all *p*-values obtained below .05 will be considered statistically significant. This means that, with the results obtained, there is only a 1 in 20 chance that the results would have been obtained if there were no difference between the groups in the population.

The investigator is free to choose whatever alpha (*p*-value) appears appropriate for a particular study. The higher the *p*-value, the easier it is to obtain statistical significance, but the chance is greater that a type I error has been made. Consequently, with a high alpha the investigator may claim that the treatment had an effect when, in reality, there was no effect, and the differences that resulted were actually seen as a result of chance.

Type II Errors

Type II errors occur when the investigator did not find a difference between two or more groups that was actually present. The frequency of type II errors is inversely related to the frequency of type I errors and is thus inversely related to the alpha set by the investigator. The lower the alpha, the higher the probability of a type II error. The chance of making a type II error is referred to as beta. Power is the inverse of beta and is the chance of finding a difference that is actually there. Several factors determine the power of significance tests, one being sample size.[4]

Because of the influence that the alpha level has on both type I and type II errors, this value must be set carefully. An alpha that is too high may result in statistical significance when no actual difference was present in the data. However, if the alpha is set too low, real differences in the data may remain undetected.

The investigator should determine the appropriate alpha based on the consequences of making either a type I or II error. If the investigator is examining a new treatment for a fatal disease, they may wish to find anything that might actually help patients. Consequently, the investigator would rather make a type I error—say something worked when it really did not—and set the alpha at .05 or even .10. However, if a new drug for a minor disease had potentially harmful side effects, the investigator would rather say that it was ineffective, even if the drug actually was effective (type II error). In this case, the investigator would set the alpha at a lower level (.01). Other reasons for adjusting the alpha related to the specific type of statistical analysis are beyond the scope of this paper.

Conclusion

Formulating and testing hypotheses is an essential component of statistical analysis. There are, in general, four steps to hypothesis testing. First, the null and alternative or research hypotheses are defined. The null hypothesis states that there is no difference between, for example, the treatment and control groups. The research hypothesis is the opposite and states that there is a difference between the treatment and control groups.

Second, the significance level or alpha level is defined. The alpha level determines what percentage of the time the investigator is willing to falsely reject H_0 or incorrectly conclude that the treatment had an effect. The alpha is also used to determine the critical value that defines the threshold the obtained test statistic must exceed and thus fall within the critical region to obtain statistical significance.

Third, the test statistic is computed, and fourth, the obtained test statistic is compared with the critical value. If the obtained test statistic falls within the critical region, the investigator rejects the null hypothesis and concludes that the treatment did have an effect and the research hypothesis is plausible. If the obtained test statistic does not fall within the critical region, the null hypothesis is not rejected, and the investigator concludes that there is no treatment effect. The *p*-value associated with the test statistic is the probability of obtaining a result as extreme as or more extreme than the one that was actually observed, given the null hypothesis is true. Smaller *p*-values (ie, < .05) suggest that there is evidence to conclude that the treatment had an effect.

Regardless of what decision is made, there is some probability for error. The alpha level determines the amount of error the investigator is willing to accept. This kind of error, type I error, occurs when the investigator rejects H_0 when, in fact, there was no treatment effect. The inverse of a type I error, a type II error or beta, occurs when the investigator retains H_0 when there was a treatment effect. The consequences of such errors should be determined before setting the alpha level.

Continued on page 153

Basics of Research Part 14: Hypothesis Testing
Continued from page 110

References

1. Thompson CB, Panacek EA. Basics of research part 3: Research study designs: Experimental and quasi-experimental. Air Med J 2006;25:242-6.
2. Thompson CB, Panacek EA. Basics of research part 1: Clinical research and critical care transport: How to get started. Air Med J 2006;25:107-11.
3. Diekhoff GM. Basic statistics for the social and behavioral sciences. Upper Saddle River, NJ: Prentice Hall; 1996.
4. Cohen J. Statistical power analysis for the behavioral science. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

Shane Allua, PhD, is a senior consultant for a global business consulting company located in Bethesda, Maryland. Cheryl Bagley Thompson, PhD, RN, is an associate professor and assistant dean of informatics and learning technologies at the University of Nebraska Medical Center College of Nursing in Omaha. She can be reached at cbthompson@unmc.edu.

1067-991X/$36.00

doi:10.1067/j.amj.200x.xx.xxx