

An Examination of Student Mistakes in Setting Up Hypothesis Testing Problems

By W. Conway Link
Department of Mathematics
Louisiana State University in Shreveport

Abstract

In setting up a typical hypothesis testing problem, the opportunities for mistakes are many. Mistakes made in the statement of the null and or alternative hypothesis can have a major impact in the interpretation of the results. A detailed analysis of student responses on such problems will be presented.

An Examination of Student Mistakes in Setting Up Hypothesis Testing Problems

By W. Conway Link
Department of Mathematics
Louisiana State University in Shreveport

Hypothesis testing is one of the more difficult topics encountered in an elementary statistics course. Textbook authors have various ways of explaining the concepts, sometimes incorporating a multi-step approach which requires the student to supply one or more of the following: the null and alternative hypothesis, the critical value of the test statistic necessary to reject the null hypothesis, the observed value of the test statistic, the decision about the null hypothesis, and the p-value. Such a multi-step approach helps guide the student to the solution in an organized manner.

For many years a six-part procedure for setting up hypothesis testing problems has been presented to statistics students at LSU-Shreveport. The first part in this procedure is the statement of the null and alternative hypothesis. Students look for key words and phrases such as “less than”, “decreased”, “reduces”, “greater than”, “increased”, “improved”, and “is different from”, as guides in stating the null and alternative hypothesis.

In the second part, the critical value of the test statistic necessary to reject the null hypothesis is asked for, which requires that the student recognize the appropriate test statistic, locate the correct tabled value based on the stated level of significance, and supply the correct sign, since the relationship between the direction (less than, greater than, not equal to) of the alternative hypothesis is related to the value of the test statistic (negative, positive, or plus/minus).

As a space saving device, many textbook examples and exam problems provide the hypothesized value of the population parameter and the necessary estimates of the appropriate parameters, rather than actual data. In worked-out solutions to example problems, such as those found in *Introduction to Probability and Statistics*, 10th ed. by Mendenhall, Beaver, and Beaver, it is quite common to give the formula for the test statistic followed by the plugged-in values, and, finally, the calculated observed value of the test statistic.

In the third part of the multi-step procedure, an alternate approach, indicative of a higher level of understanding of hypothesis testing, requires the student to construct a chain of probability statements about obtaining the observed value of the sample statistic or one

more extreme given the null hypothesis is true. For example, if the alternate hypothesis is $\mu > 3.7$ with $\bar{x} = 3.9$, $s = 0.3$ with $n = 9$, the chain would be: $P(\bar{x} \geq 3.9) = P(t \geq [(3.9 - 3.7) \sqrt{9}] \div 0.3) = P(t \geq 2)$.

While completing the chain of probability statements, the student calculates the observed value of the test statistic. So, in the fourth step in the six-step process, the student re-writes this value (in the above example, $t = 2$).

In the fifth step, the observed value of the test statistic is compared with the critical value necessary to reject the null hypothesis, resulting in a decision about the null hypothesis.

Finally, the p-value is either read from a table, or is displayed on a graphing calculator screen.

This six-step procedure (or any similar multi-step procedure) seems to be a reasonable way to enhance and test the student's understanding of hypothesis testing, since incorrect responses for the individual items indicates that full understanding has not been achieved. Typically, individual items on the multi-step hypothesis testing problems are graded as correct or incorrect without taking into consideration the frequency with which each is missed, and the implications of making an error at any point in the process.

To estimate the frequency with which the six individual items are answered incorrectly, the responses to the hypothesis testing problems on six recent in-class tests taken by 295 students were examined. Of the six tests, five were finals exams. Two courses were represented -- Applied Statistics, a requirement for Biological Sciences majors, has only College Algebra as a prerequisite, and Elementary Statistics, which has a calculus prerequisite. Included on these exams were hypothesis tests about a single population proportion ($<$ and $>$), hypothesis tests about a single population mean ($<$ small n , $>$ large n , and \neq large n), hypothesis tests about the equality of population means (both small n and large n), and the equality of population proportions. Typically, there were three or four hypothesis testing problems on each exam, but the hypotheses to be tested varied from exam to exam.

Results

A proper statement of the null and alternative hypothesis is a critical component in successfully completing and understanding the hypothesis-testing concept since much of what follows in the multi-step procedure is based on the correctness of this item. To complicate the matter, there are several ways of responding incorrectly, each of which is indicative of a different level of understanding (or non-understanding) of the hypothesis-testing concept. For the purpose of this paper, four categories for errors and a non-response category were established: (a) statement about the wrong population parameter; (b) statement about a sample statistic (such as \bar{x} instead of μ), (c) correct statement about the population parameter, but containing an error in the hypothesized value, sign,

or direction of inequality; and (4) meaningless statement.

In examining the $n = 295$ test papers, it was found that 12.6% of the students gave statements about the wrong population parameter, 5.8% made statements about a sample statistic such as \bar{x} instead of μ , 12.6% correctly identified the population parameter but made an error in the hypothesized value, sign, or direction, and 0.7% gave a meaningless response. Sixty-seven percent of the responses were correct. Non-response on this item was low – 1.4%.

Certain population parameters about which hypotheses were tested were more troublesome than others. When a test about the equality of population proportions was called for, 50% of the $n = 26$ responses were incorrect. Students incorrectly tested a hypothesis about sample statistics 19.2% of the time and about the wrong population parameter 23.1% of the time.

Giving the correct value of the test statistic necessary to reject the null hypothesis proved to be somewhat more difficult – 46.9% of the responses on this item were correct. Incorrect responses were assigned to three broad categories – (a) correct test statistic but incorrect value (39.8%), (b) wrong test statistic (4.1%), and (c) meaningless response (3.1%). The percent non-response was 6.1%.

A different two-tailed test, a test about the equality of population means for large samples, gave students the most difficulty when they were asked to find the critical value of the test statistic – only 29.2% ($n = 48$) responded correctly. The wrong value was given in 39.6% of the responses and an incorrect test statistic was used 12.5% of the time. Students were more successful with the hypothesis test about a single population proportion ($>$) – 71.8% responded to this item correctly ($n = 39$).

Constructing a probability statement about the observed value of the sample statistic proved to be the most difficult – only 26.2% responded correctly. A slightly larger percentage, 26.9%, started with a formula, rather than the requested probability statement. An incorrect observed value or wrong inequality direction was given 17.3% of the time, an incorrect test statistic was used 4.4% of the time, a probability statement about a population parameter rather than a sample statistic was given 11.9% of the time, and 3.1% of the responses were meaningless. Non-response on this item was 10.2%.

Two-tailed tests again proved to be the most troublesome when students were asked to make a chain of probability statements about what was observed. Only 12.5% responded correctly for the test about the equality of population means for large sample sizes ($n = 48$), and only 15.4% successfully responded for the test about the equality of population proportions ($n = 26$).

Students had an easier time obtaining the correct observed value of the test statistic –

65.1% of the responses to this item were correct, 25.8% were incorrect, and 0.3% were meaningless. Non-response was 7.8% on this item. The two tailed test about the equality of population proportions proved to be the most difficult – only 33.3% of these calculations were correctly done.

Deciding what to do with the null hypothesis proved to be the easiest for the students. Overall, a correct decision was made 78.6% of the time, an incorrect decision 16.3% of the time, and a meaningless statement 0.3% of the time. Non response was low – 4.7%. Again, students found the most difficulty with a two-tailed test, this time it was a test about the equality of population means for large sample sizes – 58.3% of the calculations were correct (n = 48).

Finally, 55.9% responded correctly when giving the p-value, 26.8% gave an incorrect answer, and 5.1% had a meaningless answer. Non-response on this item was the highest among all items – 12.2%. When specific hypothesis tests were taken into account, students found the determination of a p-value the most difficult when testing a hypothesis about a single population proportion (greater than direction) – only 43.6% of the responses were correct (n = 39). However, the next-to-lowest correct response percentage (47.1%) did involve a two-tailed test – a test about the equality of population means for small sample sizes (n = 17).

Discussion

1. Correct statements about the null and alternative hypothesis are important because they set the stage for much of what follows in the multi-step hypothesis testing procedure. A statement about a sample statistic (e.g. \bar{x} instead of μ) or the wrong population parameter (e.g. μ instead of P) is a clear indication that the student lacks full understanding of the hypothesis-testing concept. Those who correctly identify the population parameter involved, but either give the wrong hypothesized value of that parameter or an incorrect direction of the inequality, increase the risk of making an error in subsequent parts of the multi-step process. For example, if the problem requires a one-tailed test, and by mistake, the student indicates a two-tailed test, the opportunity for errors in the critical value of the test statistic, the probability statement, and the decision about the null hypothesis will be greater, since each is dependent on the alternative hypothesis being correctly stated.

2. Since a decision about the null hypothesis is based on the comparison of the observed and critical values of the test statistic, an incorrect response here may affect the decision about the null hypothesis. In this study, 39.7% of the responses to this item were incorrect (wrong value or sign), which suggests that even though the appropriate test statistic was identified, there was some difficulty in locating its correct tabled value based on the stated level of significance or applying the correct sign possibly due to an error in the statement of the alternative hypothesis.

3. The low percentage of correct response (26.2%) to the probability statement item can be attributed to one or more of the following: (a) failure to understand the overall goal - that of finding the probability of obtaining the observed value of the sample statistic or one more extreme given that the null hypothesis is true; (b) failure to understand the connection between the probability statement about the observed value of the sample statistic and the probability statement about the appropriate test statistic; and (c) failure to use the correct formula. It should be noted that each student was provided with a formula – only sheet, and it was up to the student to choose the formula appropriate to the application.

4. Calculating the observed value of the test statistic is actually a part of and the last step in the probability statement item. The low percent (26.1%) correct response for the probability statement item and the much higher percent (65.1%) correct response for this item supports the belief that while students can correctly substitute values into a formula selected from a formula sheet, they do not have full understanding of the hypothesis testing process.

5. The item receiving the highest percent of correct responses was that regarding the decision about the null hypothesis. Since there was nothing to calculate or look up in a table (only a “reject” or “don’t reject” response was required), it is quite likely that lucky guesses helped boost the percentage of correct responses in this category. The best evidence for this is that in several instances, a correct response to this item was unrelated to the student’s observed and critical values of the test statistic. For example, if $z = 1.645$ or greater was required to reject the null hypothesis, it was not uncommon for a student to have an observed value less than 1.645, yet still reject the null hypothesis.

6. Although most of the students taking the exams had access to a graphing calculator such as the TI-83 which can display p-values on the screen, 31.9 % of the responses to this item were either incorrect or meaningless. Errors by those using such a calculator could be due in part to one or more of the following: (a) selecting a two-tailed test on the menu when a one-tailed test is required or vice versa; (b) selection of the wrong tail from the menu when using a one-tailed test; (c) incorrect entry of one or more observations in the list(s); (d) incorrect entry of one or more sample statistics; or (e) incorrect selection from the “Tests” menu. No record was kept of which student used a graphing calculator, so it was not possible to determine in which category the incorrect responses fell.