For Unequal Samples of Skewed Data - Which t-test: Unequal or Equal Variances?

Avraham Wein[1], James Schmeidler[2]
[1]Yeshiva College, 500 West 185th Street, New York, NY 10033
[2]Icahn School of Medicine at Mount Sinai, 1428 Madison Ave, New York, NY 10029

Abstract
Student's t-test is appropriate for testing the difference of means for normally distributed data with equal variances, as is an approximate t-test for unequal variances. The conventional strategy chooses between them using Levene's test for equality of sample variances. Which test should be used if both samples have the same skewed distribution; what is the effect of unbalanced sample sizes? Examples have the dichotomous Bernoulli distribution with success probability .10. The sample means have binomial distributions, facilitating evaluation of significance comparing a sample of 20 with other samples, under the null hypothesis of the same distribution. Testing significance for an outcome from each sample is simplified by using a weighted data set. For this example, Levene's test is more sensitive than either t-test, so the conventional test is the same as unequal variances. For both sample sizes 20, each two-sided test is symmetric. As the larger sample size increases, the two t-tests favor opposite sides. This bias makes each t-test inappropriate for either two- or one-sided testing. Kurtosis associated with skewedness also makes the unequal variances test liberal.

1. Choosing the Appropriate t-test

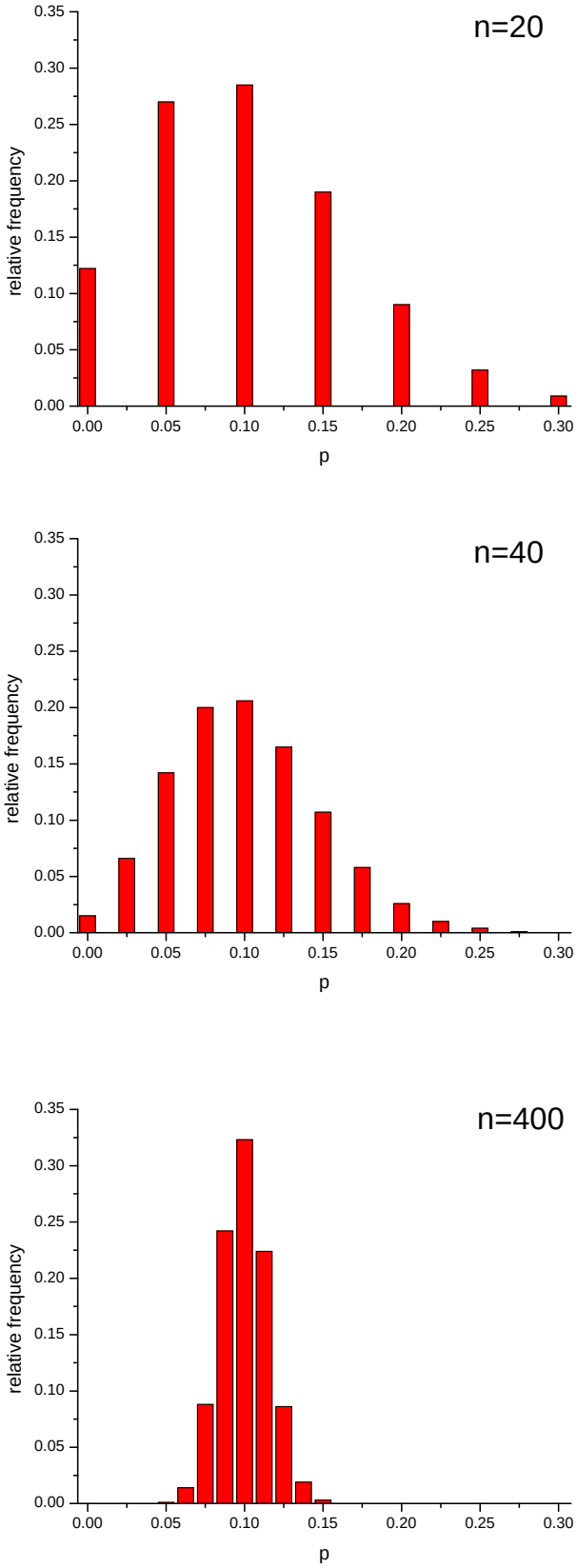
Standard statistical practice dictates the use of certain procedures if one is seeking to determine whether the means of two populations differ. If one makes the assumption that the populations possess normal distributions then several procedures are available. Student's t-test assumes equal standard deviations. If equal variances are not assumed an approximate t-test can be used. The conventional practice is to decide between the two t-tests by evaluating the significance of Levene's test for equality of variances of the samples of the populations. We are implementing all three of these tests in SPSS version 21. SPSS employs the original Levene's test (Levene, 1960). SPSS does not use the method suggested by Brown and Forsythe (Brown and Forsythe, 1974).

What happens if the two samples come from the same distribution, which is not a normal distribution? If the sample sizes are equal, no substantial problems are introduced and Student's t-test remains a good option. When sample sizes differ and the distribution has substantial positive kurtosis, for both t-tests the probability of rejecting the null hypothesis by chance is larger than the nominal significance level. In an introductory discussion of non-normality the effects of kurtosis are sometimes mentioned but not the effects of skewedness. In this paper we demonstrate how the adverse effects of skewedness can be presented at an elementary level accessible to an introductory course.

2. Illustrating the Effects of Skewedness

In order to illustrate the effects of skewedness at an elementary level, we begin by selecting a particularly simple skewed distribution. When the data have a dichotomous Bernoulli distribution with a probability .1 of success, the sample mean has a binomial distribution. To show the effect of unequal sample sizes we compare a sample of 20 to samples of 20, 40 and 400. For n=400, we grouped probabilities to create a bar graph comparable to the bar graphs of the other distributions.

Figure 1. Relative frequency distributions



To compare the three procedures we find the probability of significance for each one-sided test under the null hypothesis of equal means. We can evaluate this probability by examining each possible pair of a result from one sample and a result from the other sample, and adding up the joint probabilities of those pairs for which the t-test is significant.

2.1 Evaluation of a Comparison Between Samples

It is not necessary to create all the data to test the significance of a difference between an outcome from the sample of 20 and the outcome from one of our other samples. For example, the sample of 20 has one success (proportion of successes equals .05) and the sample of 400 has 64 successes (proportion of successes equals .16). A shortcut that avoids creating a sample of 20 observations with one success and another sample of 400 observations with 64 successes is to use a data analysis instruction to weight a data set (see Figure 2).

Figure 2- SPSS Data Set for Weighted Analysis

sample	x	N
N = 20	1	1
N = 20	0	19
N = 400	1	64
N = 400	0	336

The simple data set has only four cases; the success in the sample of 20, the failure of the sample of 20, the success of the sample of 400 and the failure in the sample of 400. When we weight the success in the sample of 20 by 1, the failure of the sample of 20 by 19, the success in the sample of 400 by 64 and the failure in the sample of 400 by 336 we analyze the data set as if it had 20 observations in the first sample and 400 observations in the second sample.

2.2 Summarizing Results From Pairs of Samples

It is not necessary to evaluate probabilities of all pairs of outcomes to find the probability of a one-sided significance. For each outcome in the sample of 20, we can evaluate the significance of all outcomes from the other sample. (However, it is only necessary to evaluate outcomes in the sample of 20 that have probability greater than .0005-- outcomes of 0-8 successes.) The result will be significant only if the proportion in the other sample differs substantially from the proportion in the sample of 20. In order to find the limit of these extreme proportions, we use the “method of halving the interval.” It starts with the interval from the proportion in the sample of 20 to one end. Evaluating the significance for the proportion in the middle of that interval indicates whether that middle proportion is extreme or not extreme. If the value is significant the next interval to evaluate is the interval from the proportion in the sample of 20 to the middle proportion. If the result in the next interval is not significant, the next interval to examine is the interval from the middle proportion to the end. By successively halving the interval, we can identify the least extreme proportion in the other sample for which that -one-sided

test is significant. For this proportion in the sample of 20, the conditional probability of a significant result in the other sample is that tail of its binomial distribution. The probability of a significant one-sided test is the sum – over all outcomes of the sample of 20– of the probability of that outcome of the sample of 20 multiplied by the conditional probability of a significant result in the other sample.

In all these examples, Levene's test was more sensitive than both t-tests. Thus if either t-test was significant, Levene's test was also significant, so the conventional test was always the approximate test.

2.3 Differences Between the Student's t-test and the Approximate t-test

The numerators of Student's t-test and approximate t-test are the same: the mean of the sample of 20 minus the mean of the other sample. The denominator of each t-test is the square root of an estimate of the variance of the numerator as a weighted average of the sample variances. For Student's t-test, the variance of the larger sample primarily contributes to this weighted average, but for the approximate t-test, the primary contribution is the variance of the smaller sample.

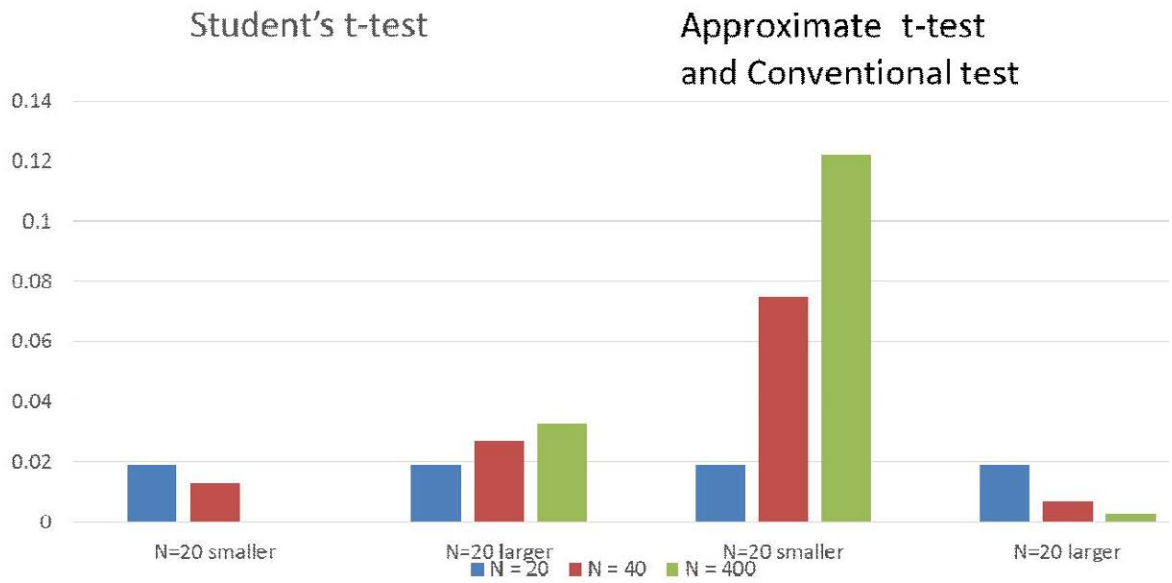
All three two-sided tests (Student's t-test, the approximate t-test, and the conventional test) are symmetric when both sample sizes are equal to 20. However for unequal sample sizes, each two-sided test is composed of one-sided tests with discrepant differences between them for Student's t-test and the approximate t-test. The discrepancy between Student's t-test and the approximate t-test reflects the differences in distributions as the sample size increases. The expected value of the sample mean is always .10, it is not changed by sample size changes. In contrast, the standard deviation of the sample declines as the sample size increases from .067 for n=20, to .047 for n=40, to .015 for n=400, and eventually to zero (see Figure 1). The skewedness and the correlation of the sample mean and the sample variance also decline to zero. A directional difference in the sample means is associated with the sample variances, which affect the denominators of the t-tests, and eventually the t-tests.

When the mean of the sample of 20 is smaller than the mean of a larger sample, the sample variance of the sample of 20 is also relatively small, due to its correlation with the sample mean. This is the primary contributor to the denominator of the approximate t-test, which is thus also relatively small compared to the denominator of Student's t-test. That makes the approximate test stronger than Student's t-test. In contrast, when the sample mean of the sample of 20 is larger than the mean of the other sample, the opposite occurs. The sample variance of the sample of 20 is relatively large, as is the denominator of the approximate t-test. Thus Student's t-test is stronger than the approximate t-test in this case.

2.4 Results of Numerical Evaluations

Figure 3- Probabilities of one-sided significance (α = .025)

Figure 3 presents the probabilities of significant results of one-sided tests.



These examples illustrate several important effects of skewed distributions. (Unfortunately, since in these examples Levene's test is more sensitive than both t-tests, the conventional test was always the approximate t-test. Therefore this example does not shed any light on the use of the conventional procedure.) For unequal sample sizes, Student's t-test and the approximate t-test favor opposite sides of their two sided tests. As sample sizes increase, the probability of rejection of a one-sided test becomes farther from the nominal value. Since a two-sided test is composed of two one-sided tests the discrepancies of the one-sided tests from their nominal significance levels are in opposite directions and thus partially cancel. Overall, the approximate t-test becomes more liberal than Student's t-test as the discrepancies of sample sizes increase.

3. Implications of Skewed Distributions and Unequal Sample Sizes for t-tests

3.1 t-tests are Inappropriate for Substantially Unequal Sample Sizes

These examples demonstrate how one-sided t-tests fail to achieve nominal significance levels. Thus one-sided t-tests clearly do not perform as intended. Although a two-sided t-test may have an approximately correct level of significance, this does not negate the problem of bias. This problem is not that false rejections will tend to be on one side but rather the unacceptably low power for one side. Moreover, one could manipulate the results of a statistical analysis by transforming the dependent variable.

3.2 Alternative Analyses

If t-tests for skewed distributions with unequal sample sizes are inappropriate, what should be done? A more general question than the difference of means is whether the distributions of the two samples differ. A variety of non-parametric tests are presented in introductory statistics. This could be an opportunity to introduce these tests if they were not previously mentioned in the course syllabus.

References

1. Levene, H. (1960), Robust tests for equality of variances. In: Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, ed. I. Olkin,. Palo Alto, CA: Stanford University Press, pp. 278-292.

2. Brown, M. B., and A. B. Forsythe. (1974), “Robust tests for the equality of variances,” Journal of the American Statistical Association, 69: 364-367.