Simple Tests of Hypotheses for the Non-statistician: What They Are and Why They Can Go Bad

Arthur L. Carpenter
California Occidental Consultants

ABSTRACT

Hypothesis testing is a central component of most statistical analyses. The focal point of these tests is often the significance level, but what does this value really mean and how can we effectively use it? And perhaps more importantly, what are the pitfalls and dangers in its interpretation? As we conduct statistical tests of hypotheses, are there other things that we should be doing or looking out for that will aid us in our decision making?

After years of academic study, professional statisticians will spend additional years gaining a practical knowledge and experience so that they can correctly conduct and correctly interpret the results of statistical tests. You cannot gain these insights over night, however, this tutorial provides a basic introduction to the concepts of hypothesis testing, as well as, what you should look for and look out for, while conducting statistical tests of hypotheses.

KEYWORDS

significance level, hypothesis test, T-test, Type I error, Type II error, alpha level

INTRODUCTION TO STATISTICAL TESTING

The point of collecting data is to be able to answer questions. Statistically these questions MUST be framed as 'hypotheses' and these hypotheses, in turn, can be tested. The results of the tests are interpreted and conclusions are drawn about the original questions. The process seems very straight forward. SAS provides all the processing capability and all we have to provide are the data. The problem, of course, is that the world of testing is not quite that simple and things can go wrong to the point that it becomes difficult, if not impossible, to draw valid conclusions.

The way that we collect our data and the statement of our questions to be answered, suggest certain 'tests of hypotheses'. These tests are based on rigorously developed and mathematically proven techniques, however to be valid they must be appropriate for the data and the situation. This appropriateness defines a set of assumptions, and unless these assumptions are met, distortions are introduced in what are often unknown and perhaps even unknowable ways. The danger of course is that these distortions can cause us to draw incorrect conclusions about our initial hypotheses.

Taking a Sample

Very often we need to draw conclusions about a population that we are unable to fully measure. For instance if we wanted to compare the weights of male and female teenagers, we simply cannot weigh all teenagers in the United States (if for no other reason than they don't hold still long enough). Fortunately we can develop statistical estimates about the overall population by taking a subset. This subset of the population is known as a sample, and if it is taken correctly the estimates based on the sample can be applied to the population.

By implication as the size of our sample increases, the reliability of our estimates that are based on the data will be more accurate. Of course increasing the sample size is often expensive and past some point, can be subject to the laws of diminishing returns.

A Typical Value - Using the Mean of a Sample

One of the types of information (statistics) that we often want to determine is a measure of centrality. One way to think of centrality is to determine what a typical value of the sample might be. What is the weight of a typical teenage male? Aside from the fact that there probably are no typical male teenagers, there are several statistics that would help us make this estimate. The MEDIAN is the value for which 50 percent of those in sample had values above (and below) the median. Another measure of centrality is the MEAN. There are several ways, formulas, that can be used to calculate the mean. The most common of these is known as the arithmetic mean and it is calculated by dividing the sum of the values by the number of values.

The formulas for the mean and the median will result in essentially the same number when the distribution of values is fairly symmetrical. Because it has a number of desirable statistical properties, the mean is used for a great many of the common statistical tests.

Measuring Dispersion - Using the Variance

If everyone in the population had exactly the same weight, then so too would everyone in the sample, and consequently the mean would be an exact measure of a ‘typical’ value. Since the weights of individuals vary, it is likely that the calculated mean of a data sample would be different for different subsets of the population. If all of the individuals have very similar weights, then they will be close to the mean, however if the weights fluctuate a great deal, then fewer individuals will have a value close to the ‘typical’ value. Thus when we measure the mean we also need to know how good the estimate is. The estimate of the mean is more precise if the weights are in closer agreement.

The dispersion in the data, and indirectly the precision of the estimate, is measured by the variance. This statistic measures the variability or dispersion in the data. An estimate of this variability will help us to assess how much we can trust our estimate of the mean.

Another common measure of dispersion is the standard deviation which is the square root of the variance. The standard deviation is often used as it has the same units as the mean.

Distribution of the Mean

If you were to take a number of samples, and you calculated an estimate of the mean each time, the mean values will themselves have a distribution. According to the Central Limit Theorem the resulting distribution of values can be described using the normal distribution. This is also sometimes referred to as the ‘Bell Shaped Curve’.

This property of the mean and the characteristics of the Normal distribution combine to give the statistician a number of very nice properties that are used in many of the parametric statistical tests that are commonly used.

Hypotheses

Simply describing a sample (or a population) is rarely sufficient. Usually we need to ask questions of the data *e.g.* is the mean weight of males different from the mean weight of females. In order to ask this question statistically we need to rephrase it as an hypothesis. It is this hypothesis that becomes central to the testing that is discussed below.

The hypotheses to be tested determine everything from the experimental design to the types of statistical analyses that are to be utilized. In a well designed study the researcher first determines the question of interest, phrases it as an hypothesis, and then determines how the data are to be collected and analyzed.

Unfortunately too many researchers use the PARC method (**P**lanning **A**fter **R**esearch **C**ompleted) for designing their experiment. They first collect the data and then try to determine what questions can be asked. The utility of this method is best described by spelling its name backwards.

The NULL hypothesis (H_o) is best stated in terms of equalities *e.g.* things are not different, for instance:

The mean weight of the males in the study is not different from that of the females.

The Alternative hypothesis (H_A) becomes the secondary hypothesis, if the null hypothesis is disproved.:

The mean weight of the males in the study is greater than that of the females.

The statistical tests calculate significance levels, and it is these significance levels which are used as a measure of the validity of the hypotheses. The significance level, which is also called the ‘alpha level’, is often indicated with Greek letter alpha, α. If the significance level is small then it is unlikely that the null hypothesis can be accepted.

Since our statistics are based on a sample, we cannot be assured that our particular sample fully represents that of the population. This means that we may reject a null hypothesis based on our particular sample, when in fact the hypothesis was true. In fact the significance level tells us the probability of making that mistake. By tradition most researchers will reject the null hypothesis when the significance level is below .05 (they are willing to accept a 5% risk of making the mistake of rejecting a true null hypothesis).

Actually when testing hypotheses you can make one of two kinds of errors. These are known as Type I & Type II errors.

		Actual Status of the Null Hypothesis	
		H_o: Is True	H_o: Is False
Outcome of the test	Test is Significant (Reject the H_o)	Commit a **Type I** Error. The probability of making this mistake is α.	Correct Decision
	Test is not Significant (Fail to reject the H_o)	Correct Decision	Commit a **Type II** Error. The probability of making this mistake is β.

When the null hypothesis is true and yet the test is significant (low alpha level), we will incorrectly reject the null hypothesis. This results in our making what is known as a **Type I** error.

When the test is not significant and we accept the Null hypothesis, we run the risk of making a **Type II** error. A measure of the risk of making a Type II error is β, and the POWER of the test is $1-\beta$.

Obviously we want to design our tests to minimize both α and β, thereby minimizing our risk of making either type of error. Increasing our sample size and decreasing our variances gives us the most control for the minimization of these two measures.

It turns out that α is a fairly straight forward calculation for most tests, while β is fairly difficult. Consequently phrasing the null hypothesis as an equality, when in fact we want to show the opposite *i.e.* reject the null hypothesis, allows us to make the determination using just the alpha level.

TESTING AN HYPOTHESIS USING A T-TEST

One of the simpler and more common types of hypothesis test involves the comparison of typical values from two groups or categories. Generally in this type of test we want to determine if the mean value from one group is the same or different from the mean of a second group. T-tests are used extensively to make these types of comparisons. There is more than one type of T-test and in the discussion that follows we will cover the most commonly used; the two sample T-test.

Assumptions of the Test

As was mentioned earlier, the mathematics of any given statistical test requires a set of assumptions. For the T-test the primary assumptions are:

- Values are measured on the ratio scale
  We need to be able to calculate the mean and variance accurately

- Errors are distributed normally
  A number of the internal calculations depend on the assumption of normality

- The dispersion (variance) is the same for both samples
  The variance tells us how accurately we are measuring the mean. Most calculations assume that both means have been measured with the same level of accuracy.

- Samples are taken independently
  Correctly collecting the sample can be a science in-and-of itself. There are any number of opportunities to violate this assumption.

Conducting a Two Sample T-test

We are conducting a small clinical study and have collected weight information on our patients. We would like to determine if among our patients there is a difference in weight between the male and the female patients. The null hypothesis might be stated as :

The mean weight of male study patients is not different from that of the female study patients.

Obviously we need to look at the means and the variances of the two genders separately. We could do this in procedures such as MEANS, SUMMARY, and UNIVARIATE (among others), however PROC TTEST will do this for us with the advantage of calculating the significance levels as well.

```
title1 'Ho: Average Weight of the Males and Females are the same';
proc ttest data=sasclass.clinics;
   class sex; ❶
   var wt; ❷
   run;
```

❶ The variable that separates the two categories (classes) must be specified.

❷ The analysis variable is specified in the VAR statement.

❸ This section of the report provides the summary statistics for analysis data. These include the means of the two groups.

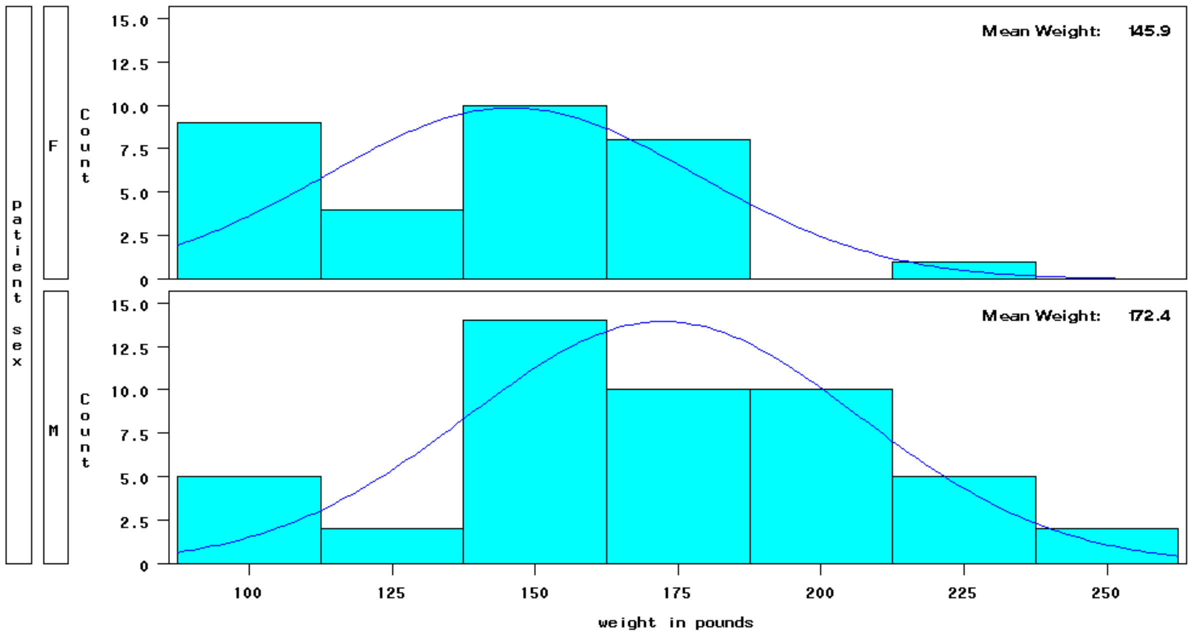
❹ This is significance level, but for a different null hypothesis. One of the assumptions of the T-test is that of equality of variances. Departure from this assumption can have severe consequences so we MUST first check to see if this assumption has been met. The null hypothesis for this test is that the variances of the two groups are equal. The high probability level (much greater than .05) suggests that we have no reason to doubt that this assumption has been met.

❺ Now that we have satisfied ourselves that we don't need to worry about unequal variances, we can check the original null hypothesis. This significance level is quite low (much less than .05), so we would reject the null hypothesis. In our study the mean weight of the males is not equal to the mean weight of the females.

```
                    Ho: Average Weight of the Males and Females are the same

                                      The TTEST Procedure

                                          Statistics

                                  Lower CL            Upper CL  Lower CL            Upper CL
Variable  SEX                N        Mean      Mean      Mean   Std Dev   Std Dev   Std Dev   Std Err


WT        F                 32      134.19    145.88    157.56    25.988    32.415    43.096    5.7303 ❸
WT        M                 48      162.39    172.38    182.36    28.633    34.395    43.082    4.9645
WT        Diff (1-2)                -41.78     -26.5    -11.22    29.073    33.622    39.871    7.6732


                                            T-Tests

             Variable    Method           Variances      DF    t Value    Pr > |t|


             WT          Pooled           Equal          78      -3.45      0.0009 ❺
             WT          Satterthwaite    Unequal      69.3      -3.50      0.0008


                                     Equality of Variances

                 Variable    Method      Num DF    Den DF    F Value    Pr > F


                 WT          Folded F        47        31       1.13    0.7369 ❹
```

We can generate a picture of the two distributions using PROC UNIVARIATE. The following step will generate the two histograms.

```
proc univariate data=sasclass.clinics;
   class sex;
   var wt;
   histogram /intertile=1 cfill=cyan vscale=count
              vaxislabel='Count'
              normal
              nrows=2;

   inset mean='Mean Weight: ' (5.2) / noframe position=ne
                                      height=2 font=swissxb;
   run;
```



Tests of Assumptions

In the above example of a PROC TTEST there is a secondary test of the assumption of equality of variances ❹. This particular test is called an F test and its name is taken from the probability distribution that is used to formulate the significance level. This is only one of a number of other tests that have been created to test this particular assumption, which is also known as homogeneity of variance or as homoscedasticity.

There are also tests for the assumption of normality (that the errors are distributed normally). Although not included in PROC TTEST, it is possible to test this assumption in PROC UNIVARIATE. It turns out that very few of the SAS procedures that perform hypothesis testing have the built in ability to test the assumptions associated with that particular procedure.

Violations of Assumptions

When (if) all the assumptions of a test are met, the calculated significance level will correctly reflect the risk of making a Type I error. Unfortunately life is such that the assumptions are rarely, if ever, truly met. The bottom line to the analyst, therefore, is a distortion to the significance level. Very few tests (the T-test shown above is an exception) have the ability to make corrections for violations to the assumptions. Since it is often hard to know how badly the assumptions are not met, and additionally what the effect on the significance level turns out to be, the distortion is usually unknown and probably unknowable.

A similar T-test to the one done above was conducted on a separate clinical trial. The results of this test are shown below. Notice that in this test the assumption of equality of variances has been violated ❻. Consequently the Satterthwaite's Correction is applied and the second significance test is used ❼. Both of the T-test alpha levels are less than .01 but notice that although the mean weights for the Males and Females are essentially the same, the two significance levels are quite different from each other.

```
                                        The TTEST Procedure

                                            Statistics

                                  Lower CL          Upper CL  Lower CL          Upper CL
Variable  sex              N       Mean     Mean      Mean   Std Dev  Std Dev   Std Dev  Std Err

wt        F               32     137.24    145.9    154.56    19.256   24.019    31.933    4.246
wt        M               48     158.85   172.38    185.92      38.8   46.609    58.381   6.7274
wt        Diff (1-2)              -44.3   -26.48    -8.659    33.915   39.221    46.511   8.9509


                                             T-Tests

             Variable    Method           Variances      DF     t Value    Pr > |t|

             wt          Pooled           Equal          78       -2.96      0.0041
             wt          Satterthwaite    Unequal      74.1       -3.33      0.0014 ❼


                                       Equality of Variances

                 Variable    Method       Num DF    Den DF    F Value    Pr > F

                 wt          Folded F         47        31       3.77    0.0002 ❻
```

If one stops to think about the process of testing of assumptions, one should quickly realize that the tests of the assumptions also have assumptions that can also be violated. For instance the F test used in the test of equality of variances ❹ ❼ is also susceptible to departures from the assumption of normality. This means that if the distribution of the errors were not normal, the test on the variances could be incorrect. What is one to do?

Interpretation of Test Results in a World of Violated Assumptions

In the above test ❼ we can clearly see the differences in the calculated significance levels. Essentially this difference is due to the violations of the assumptions of the test. When the assumption of equality of variance is not met, we calculate the Satterthwaite's corrected significance level (named for the statistician who created the correction). As was mentioned earlier very few tests have corrections such as the one used in the T-test. The T-test itself is the simplest form of the type of tests collectively known as the Analysis of Variance, ANOVA. No such correction exists for the ANOVA, and the statistician who comes up with one will surely be nominated for the statistician's hall of fame.

Although there are a few tests that we can conduct to help us determine the level that the assumptions are violated, it is even more important to understand how the test results are to be distorted. Before the availability of modern computing capabilities *i.e.* before SAS, these distortions were not clearly understood. The non-parametric or distribution free tests were developed as alternative tests when the assumption of normality could not be met an in the early 1970's a number of simulations were conduction on artificial data to try to assess the effects of violated assumptions.

In the simulation studies it was discovered that the distortions to the significance levels due to departures from normality were far less severe than originally thought. Heterogeneity of variance, on the other hand, can have a large effect (both in parametric and non-parametric tests).

When assumptions are violated, the distortions to the significance level become more pronounced for small sample sizes. [What is small is another completely separate question, which we are not going to cover here.] We can somewhat mitigate the distortion by increasing sample sizes and by controlling the distribution of the number of samples. Generally having the same number of samples in each class or cell is best and this has a number of nice statistical properties, however when variances are unequal, additional samples in the class with the larger variance can often be very helpful. In addition it is sometimes possible to use data transformations to control for both non-normality and heterogeneity of variances.

During the analysis of the results of the hypothesis test it may be helpful to remember that the calculated significance level is an estimate of the risk, not as an absolute value. For instance if you have a calculated alpha of .045 would you consider it to be significant at the .05 level? Well perhaps, but a small distortion in the assumptions could easily cause a true risk of .06 to be calculated as .045. Again what is one to do?

You might want to consider looking at a range of significance levels. I suggest to my clients that as a general rule if one is testing at the .05 level and the test results in a value less than .01, then it is likely that the true alpha level is less than .05. Conversely if the calculated level is over .1, then it is unlikely that the true value is less than .05. When the calculated significance level is between .01 and .10 the researcher must more carefully examine the data and its adherence to the assumptions of the test before making the final determination as to its true significance.

When Testing at the .05 level of α		
Likely to be significant - regardless of violations of assumptions	Significance of test requires a more careful review of assumptions	Likely to be insignificant - regardless of violations of assumptions
α < .01	.01 <= α <= .10	.10 < α

SUMMARY

The end result of statistical tests of hypotheses is an estimate of the risk of making an incorrect conclusion with regards to the acceptance or rejection of the null and alternative hypotheses. This level of risk is the probability of making a Type I or Type II error, and the significance (or alpha) level of the test is the risk associated with making a Type I error.

The significance levels are calculated using formulas that make various assumptions about the test and the data. When these assumptions are violated, the calculated significance value may not truly reflect the actual risk. Often the distortion is unknown or even unknowable. Consequently we must always be vigilant as we examine our data and the level to which our assumptions of the test are satisfied.

ABOUT THE AUTHOR

Art Carpenter's publications list includes four books, and numerous papers and posters presented at SUGI and other user group conferences. Art has been using SAS$^{®}$ since 1977 and has served in various leadership positions in local, regional, national, and international user groups. He is a SAS Certified Advanced Programmer^{TM}, and through California Occidental Consultants he teaches SAS courses and provides contract SAS programming support nationwide.

AUTHOR CONTACT

Arthur L. Carpenter
California Occidental Consultants
10606 Ketch Circle
Anchorage, AK 99515

(907) 865-9167
art@caloxy.com
www.caloxy.com

REFERENCES

Applied Statistics and the SAS® Programming Language by Ronald Cody and Jeffery K. Smith, 4th Edition, Cary, NC: SAS Institute Inc., 1999, 445pp. This book contains extensive examples on the use of SAS outside of the statistical context.

Common Statistical Methods for Clinical Research with SAS® Examples, 2nd Edition, by Glenn Walker, Cary, NC: SAS Institute Inc., 2002. A variety of statistical methods are examined through the use of SAS and SAS/STAT examples.

TRADEMARK INFORMATION