

Statistics to English Translation, Part 2b: Calculating Significance

Nina Zumel*

December, 2009

In the previous installment of the Statistics to English Translation, we discussed the technical meaning of the term "significant". In this installment, we look at how significance is calculated. This article will be a little more technically detailed than the last one, but our primary goal is still to help you decipher statements about significance in research papers: statements like " $(F(2, 864) = 6.6, p = 0.0014)$ ".

As in the last article, we will concentrate on situations where we want to test the difference of means. You should read that previous article first, so you are familiar with the terminology that we use in this one.

How is Significance Determined?

Generally speaking, we calculate significance by computing a *test statistic* from the data. If we assume a specific null hypothesis, then we know that this test statistic will be distributed in a certain way. We can then compute how likely it is to observe our value of the test statistic, if we assume that the null hypothesis is true.

We'll explain the use of a test statistic with our Sneetch example from the last installment.

The t-test for Difference of Means

Suppose that the test scores for both Star-Bellies and Plain-Bellies are normally distributed, with the means and standard deviations as given in the table below.

	n (number of subjects)	m (mean score)	s (standard error)
Star-Bellies	50	78	7
Plain-Bellies	40	74	8

Remember from the previous installment that we can estimate the true population means μ_1 and μ_2 as normally distributed around the empirical population means m_1 and m_2 respectively, with variances σ^2/n_1 and σ^2/n_2 . This is shown in Figure 1.

*<http://www.win-vector.com/>

Informally speaking, there is no significant difference in the two populations if the shaded overlap area in Figure 1 is large.

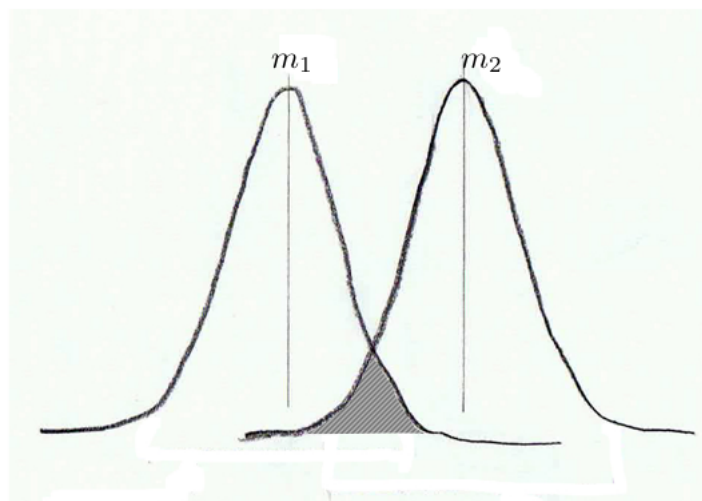


Figure 1: The estimates of the means for two populations

Calculating this area is somewhat involved. Instead, we calculate the *t-statistic*:

$$t = \frac{(m_2 - m_1)}{s_D} \quad (1)$$

where s_D is called the *pooled variance* of the two populations.

$$s_D^2 = \frac{n_1 \cdot s_1^2 + n_2 \cdot s_2^2}{n_1 + n_2 - 2} \cdot (1/n_1 + 1/n_2) \quad (2)$$

For our Sneetch example, $s_D = 1.6$, and $t = 2.499$, or the negative of that, depending on which group is Group 1. There are $50 + 40 - 2 = 88$ degrees of freedom.

If the null hypothesis is true, and the two populations are identical, then t is distributed according to *Student's distribution with $N_1 + N_2 - 2$ degrees of freedom*. Student's distribution is sort of a “stretched out” bell curve; as the degrees of freedom increase ($N_1 + N_2 \rightarrow \infty$), Student's distribution approaches the standard normal distribution, $N(0, 1)$ ¹.

In other words, if the null hypothesis is true, t should be near zero. The probability of seeing a t of a certain magnitude or greater under the null hypothesis is given by the area under the tails of Student's distribution:

¹Remember from the last installment that when you are estimating the mean of a distribution with unknown mean μ and unknown variance σ^2 , the 95% confidence interval around your estimate is $m \pm 2 \cdot \sigma / \sqrt{n}$. Intuitively speaking, Student's distribution is what you get if you calculate confidence intervals using the estimated variance s instead of the true but unknown variance σ . The distribution is stretched out compared to the normal distribution to reflect this increased uncertainty.

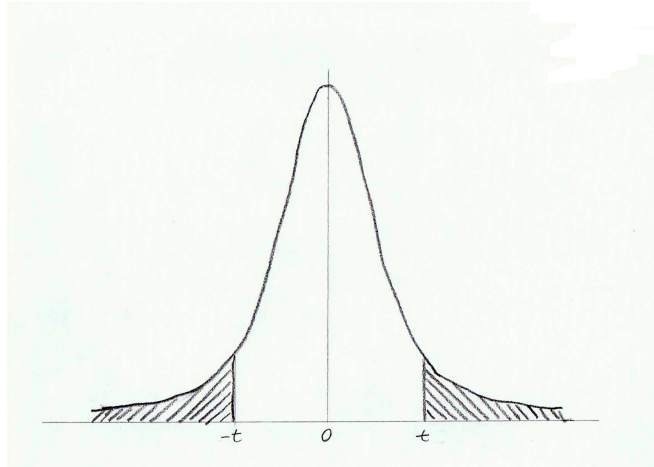


Figure 2: The area under the tails for a given t

This area is p . For the Sneetch example, $p = 0.014$.

The further out on the tails t is, the stronger the evidence that you should reject the null hypothesis. If you know for some reason that the mean of one population will be greater than or equal to the other, than you can use the *one-tailed test*:

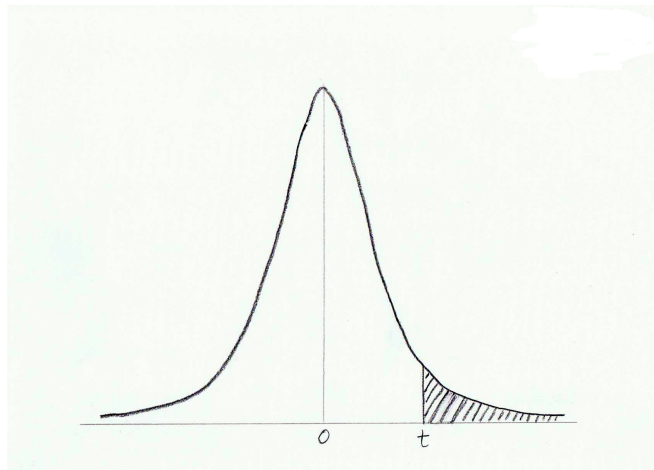


Figure 3: The one-tailed test for a given t

This test halves the p-value as compared to the two-tailed test, making a given t value twice as significant. When in doubt about which to use, the two-tailed test is more conservative against false positives².

In discussions of t-tests, you will often see statements of the form:

²In his textbook *Statistics*, Freedman tells an anecdote about a study that was published in the *Journal of the AMA*, claiming to demonstrate that cholesterol causes heart attacks. The treatment group that took a cholesterol reducing drug had “significantly fewer” heart attacks than the control group ($p \approx 0.035$). A closer reading revealed that the researchers used a one-tailed test, which is equivalent to *assuming* that the treatment group was going to have fewer heart attacks. What if the

The t-test meets the hypothesis that two means are equal if

$$|t| > t_{\alpha/2, \nu}$$

for a two-tailed test, or

$$t > t_{\alpha, \nu}$$

for a (right-sided) one-tailed test.

The quantities on the right hand side of the two equations above are called the *critical values* for a given significance level α (usually, $\alpha = 0.05$) and ν degrees of freedom. The critical values are the values for which the area of the right hand tail is equal to α .

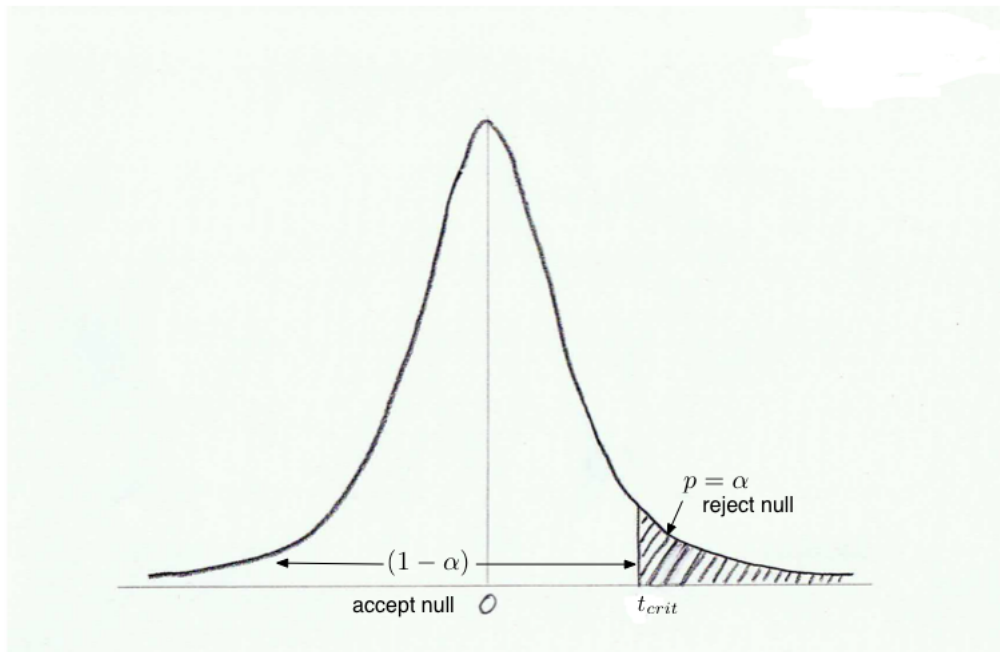


Figure 4: Critical value for a one-tailed test. Reject the null hypothesis if $t > t_{crit}$

For a two-tailed test, you must halve the area under a single tail.

drug had *increased* the risk of heart attack? The proper two-tailed significance of their results would have been $p \approx 0.07$, which is higher than *JAMA*'s strict significance threshold of 0.05. [FPP07, p. 550]

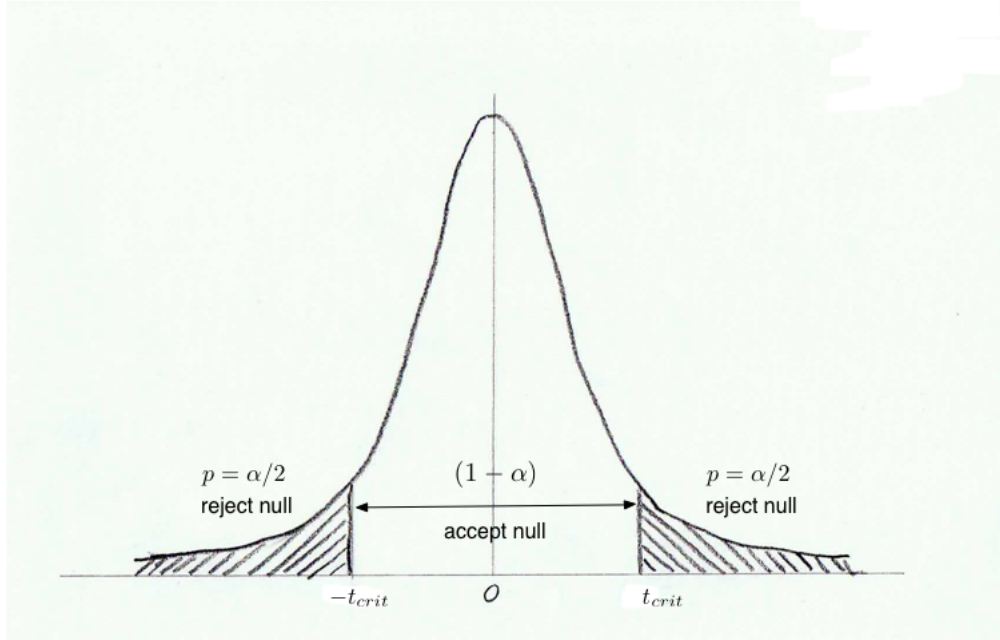


Figure 5: Critical value for a two-tailed test. Reject the null hypothesis if $|t| > t_{crit}$

This convention dates back to the time when computational resources were scarce, and researchers had to use pre-computed tables of critical values, rather than calculating p directly. Today, general statistical packages such as R or Matlab can compute the CDFs of any number of standard distributions; once you can compute the CDF, directly computing p (the area under the tails) is straightforward. Despite this, many tutorials of the t-test (and of the F-test, and other significance tests) still adhere to the convention of comparing test statistics to critical values. This tends to needlessly ritualize the whole process, and make it seem more complicated and mysterious than it actually is, at least in my opinion.

David Freedman was very much against the continued practice of using critical values, rather than reporting the actual p-value. The last chapter of Freedman, Pisani and Purves [FPP07] is worth reading for its discussion of this, and other potential pitfalls of significance tests.

Some standard packages for evaluating t-tests, F-tests, or the ANOVA also present analysis results in terms of critical values. Most of them do usually print the actual p value as well, along with the value of the test statistic and the degrees of freedom. Most researchers rightfully report the test statistics along with the actual significance levels: “we conclude that there is a significant difference in mathematical performance ($t(88) = 2.499$, $p = 0.014$)... .” Here, 88 gives the degrees of freedom, $t(88)$ is the value of the t-statistic, and p is of course the p-value.

Similar comments apply to the F-test, discussed in more detail below.

Assumptions

Strictly speaking, the t-test is only valid for normally distributed data where both populations have equal variance. However, the test is fairly robust to non-normal data [Box53]. You can verify that the sample variances are “equal enough” – that is, they could plausibly both be sampled observations from populations with the same variance, by using the *F-test*. The F-statistic

$$F = s_1^2 / s_2^2$$

is distributed according to the *F distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom*

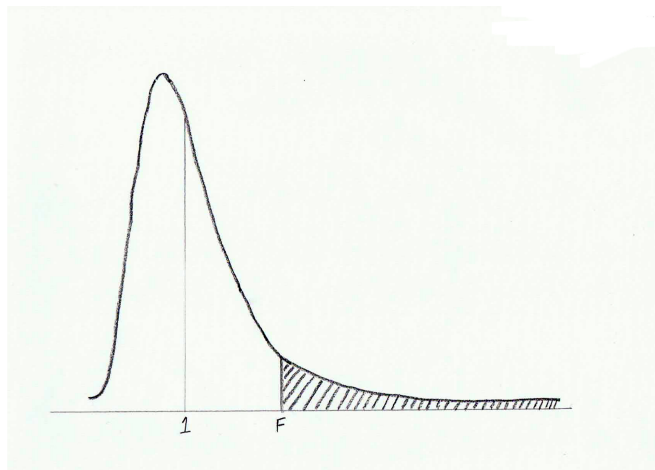


Figure 6: The F distribution

In practice, the larger variance is usually put in the numerator, so $F > 1$. The test should still be two-tailed, so you should double the area under the right-hand tail³. In this situation, you want to check if you should accept the null hypothesis (that $F \approx 1$) at a given significance level. If so, then you can go ahead and apply the t-test.

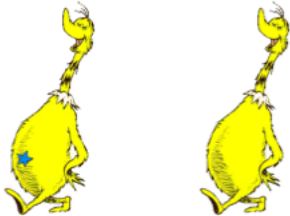
There is a variation of the t-tests for distributions of unequal variance, called Welch’s t-test [Wikc]. In this case, you are only checking if the means are equal, not that the distributions are the same.

The F-test for Analysis of Variance (ANOVA)

ANOVA is an extension of the difference of means test above to the case of more than two populations. The null hypothesis in this case is that all the sample means are equal – or more strictly, that all the treatment groups are drawn from the same population.

³The area to the right of F with (a, b) degrees of freedom is equal to the area to the left of $1/F$, with (b, a) degrees of freedom.

The simplest version of the ANOVA is the *one-way ANOVA*, where there are k *treatment groups* (populations) with n_i subjects (or repetitions, or replications) each, for a total of N subjects. Each population corresponds to a different single factor (a treatment or a condition: for example, a type of medicine, or a Star-Bellied Sneetch vs. a Plain-Bellied Sneetch vs. a Grinch). Two- or three- way ANOVAs correspond to varying two or three different factors combinatorially. For example, we could do a two-way ANOVA of Sneetch math performance by considering both the belly type and the gender of the Sneetchs.



		$m_{1,1}$	$m_{1,2}$
		$s_{1,1}$	$s_{1,2}$
		$n_{1,1}$	$n_{1,2}$
		<hr/>	
		$m_{2,1}$	$m_{2,2}$
		$s_{2,1}$	$s_{2,2}$
		$n_{2,1}$	$n_{2,2}$
		<hr/>	
	Male		
	Female		

Figure 7: Table for a Two-way ANOVA of Sneetch math performance

We will only discuss one-way ANOVA in this article, since that covers all the relevant ideas about calculating significance.

For a one-way ANOVA, we have the population means m_i and variances s_i^2 . We can also calculate the overall mean m_0 , over the entire aggregate population.

The *between-groups mean sum of squares*, which is an estimate of the *between-groups variance*, is given by

$$s_B^2 = \frac{1}{k-1} \sum_i n_i \cdot (m_i - m_0)^2 \quad (3)$$

s_B^2 is sometimes designated MS_B . It is a measure of how the population means vary with respect to the grand mean.

The *within-group mean sum of squares* is an estimate of the *within-group*

variance:

$$s_W^2 = \frac{1}{N - k} \sum_i^k \sum_j^{n_i} x_{ij} - m_i^2 \quad (4)$$

s_W^2 is sometimes designated MS_W . It is a measure of the “average population variance”.

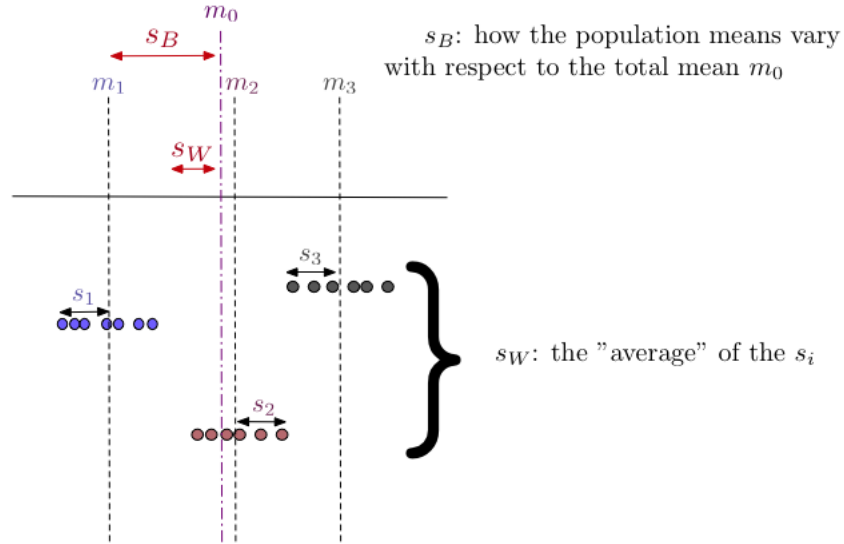


Figure 8: Within-group and between-group variance

If the null hypothesis is true, then

$$F = s_B^2 / s_W^2$$

is distributed according to the F distribution with $(k - 1, n - k)$ degrees of freedom.

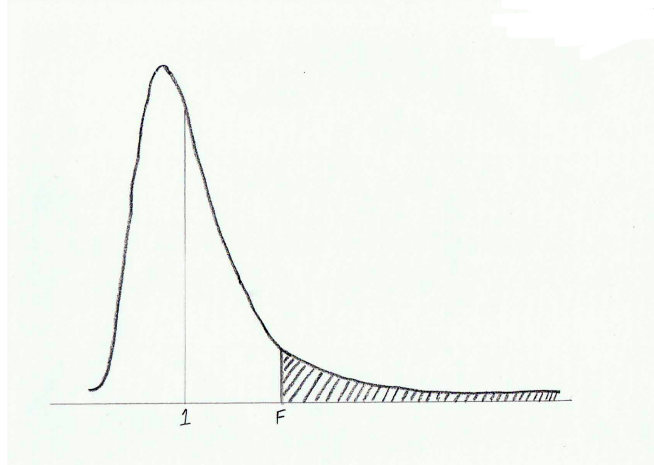


Figure 9: p-value for the one-tailed F-test

That is, under the null hypothesis, the within-group and between-group variances should be about equal: $F \approx 1$. If $F < 1$, then some of the treatment groups overlap other groups substantially, so practically speaking, one might as well accept the null hypothesis. Hence, a one-sided F test is good enough. As with the t-test, research papers usually give the value of the F statistic, the degrees of freedom, and the p-value: “($F(2, 864) = 6.6, p = 0.0014$)”. In this example, the test statistic value is 6.6, and it was evaluated against the F distribution with (2, 864) degrees of freedom, which means that $k = 3, n = 866$. The p-value is 0.0014.

Assumptions

Like the t-test, ANOVA assumes that the data is normally distributed with equal variances. According to Box [Box53], ANOVA is fairly robust to unequal variances when the population sizes are about the same, but you might want to check anyway. If all the populations are the same size (all the n_i are the same), the easiest way to check for equality of variances is an F-test of the statistic $F = s_{max}^2 / s_{min}^2$ with $n - 1$ degrees of freedom [Sac84]. In other cases, you can use Bartlett’s Test [Wika] or Levene’s Test [Wikb]. Bartlett’s test uses a test statistic that is distributed as the χ^2 distribution, and Levene’s test uses one that is distributed as the F distribution. Levene’s test does not assume normally distributed data.

If the data are not normally distributed, or have unequal variance, often they can be transformed to a form that is closer to obeying the assumptions of ANOVA. The following table of transformations is based on [Sac84, p. 517], and other sources [Hor].

Data Characteristics	Suitable Transformation
Count Data (Poisson distribution) In particular, counts of relatively rare events Data that are small whole numbers $\sigma^2 \approx k\mu$	$x' = \sqrt{x}$ Use $\sqrt{x + \epsilon}$ for some small ϵ when small values or zeros are present in the data. Sachs recommends $\epsilon = 3/8$ or 0.4 .
Percentages and proportions Binomial data Data expressed as % of control $\sigma^2 \approx k\mu(1 - \mu)$	$x' = \arcsin \sqrt{x/n}$ n is the denominator used to calculate percentages Substitute $(1/4n)$ for 0% and $(100 - 1/4n)$ for 100% . Sachs recommends $x' = \arcsin \sqrt{\frac{x+3/8}{n+3/4}}$ when 0% and 100% values are present. If the data is in the range $20\% - 80\%$, no transformation needed
Data follows a multiplicative rather than additive model Data expressed as percentage of change $\sigma \approx k\mu$	$x' = \log x$ Use $x' = \log(x + 1)$ when data has small values.

Figure 10: Table of Transformations

Jim Deacon from the University of Edinburgh lists some suggestions as well [Dea]. He also reminds us that running ANOVA on the transformed data will identify significant differences in the *transformed* data. This is *not* the same as saying there are significant differences in the original data!

Once the Null Hypothesis is Rejected

If you are able to reject the ANOVA null hypothesis, you will usually want to know which population means are significantly different from the rest. Often, in fact, you are primarily interested in which population had the highest mean. For example, if you are comparing the efficacy of a new medicine A against existing medicines B and C, you are probably not too concerned about whether B and C perform significantly differently from each other, only about whether A is significantly better than both.

If all you care about is whether the highest mean is significantly higher than the others, you can simply test where the statistic

$$(m_1 - m_2) / (s_W^2 \frac{n_1 + n_2}{n_1 \cdot n_2})$$

falls on the Student-t distribution with $n - k$ degrees of freedom. Here, s_W^2 is the within-group variance, as calculated in Equation 4, m_1 and m_2 are the highest and second highest population means, n is the total number of samples ($n = \sum n_i$), and k is the number of treatment groups.

This test is usually written

$$m_1 - m_2 > t_{(n-k, \alpha/2)} \cdot \sqrt{s_W^2 \cdot \frac{n_1 + n_2}{n_1 \cdot n_2}} = LSD_{(1,2)}$$

where $t_{(n-k, \alpha/2)}$ is the (two-sided) critical value for significance level α and $n - k$ is the number of degrees of freedom to use. This quantity is called the *least significant difference (LSD)* between the highest and second highest means, and the test is usually called the *LSD test*.

If you want to test all the population differences $m_i - m_j$ for significance, (or test the highest value against all of the others explicitly) then you need to take some care with the LSD test. Remember that a significance level of α means that with probability α you will make a false positive error. To test all possible population differences is $K = (k \text{ choose } 2)$ comparisons, or $K = k - 1$ comparisons, if you sort all the means in descending order and compare adjacent ones. Testing the highest mean against all the lower values is also $K = k - 1$ comparisons. This means you have a $K \cdot \alpha$ probability of making a false positive error. So if you want the overall significance level to be α , each individual comparison should use a stricter significance threshold $p \leq \alpha/K$.

A preferred way to compare multiple means for significance (once the ANOVA null hypothesis has been rejected) is to use a *multiple range test* [Dea] or *Tukey's method* [oST06], rather than the LSD test. Tukey's method tests all pairwise comparison simultaneously, and the multiple range test starts with the broadest range (the highest and the lowest means), and works its way in until significance is lost.

Conclusion

We've skimmed over many complications in this discussion. Hopefully, though, what we have gone over is enough to demystify much of the statistical discussion in research papers. Perhaps, it will demystify the output of standard ANOVA and t-test packages for you, as well.

Chong-ho Yu's site [hY] gives a brief discussion of some of the issues that I've skimmed over. It also lists a few common non-parametric tests. These are tests that do not make assumptions about how the data is distributed, and so they may be more appropriate for data that is very non-normal, or for discrete data. They tend to have less power than parametric tests (that is, they have a lower true positive rate); so if the data is at all normal-like, parametric tests are preferred.

Significance tests are used in other applications beyond testing the difference in means or variances. They are used for testing whether events follow an expected distribution, for testing if there is a correlation between two variables, and for evaluating the coefficients of a regression analysis. We hope to cover some of these applications in future installments of this series.

References

- [Box53] G.E.P. Box, *Non-normality and tests on variances*, Biometrika **40** (1953), no. 3/4, 318–335.
- [Dea] Jim Deacon, *A multiple range test for comparing means in an analysis of variance*, <http://www.biology.ed.ac.uk/research/groups/jdeacon/statistics/tress7.html> ([link](#)).
- [FPP07] David Freedman, Robert Pisani, and Roger Purves, *Statistics*, 4th ed., W. W. Norton & Company, New York, 2007.
- [Hor] Rich Horsley, *Transformations*, <http://www.ndsu.nodak.edu/ndsu/horsley/Transfrm.pdf>, Class notes, Plant Sciences 724, North Dakota State University.
- [hY] Chong ho Yu, *Parametric tests*, http://www.creative-wisdom.com/teaching/WBI/parametric_test.shtml ([link](#)).
- [oST06] National Institute of Standards and Technology, *Tukey's method*, NIST/SEMATECH e-Handbook of Statistical Methods, 2006, <http://itl.nist.gov/div898/handbook/prc/section4/prc471.htm> ([link](#)).
- [Sac84] Lothar Sachs, *Applied statistics: A handbook of techniques*, 2nd ed., Springer-Verlag, New York, 1984.
- [Wika] Wikipedia, *Bartlett's test*, http://en.wikipedia.org/wiki/Bartlett's_test.
- [Wikb] ———, *Levene's test*, http://en.wikipedia.org/wiki/Levene's_test.
- [Wikc] ———, *Welch's t test*, http://en.wikipedia.org/wiki/Welch's_t_test.