
Statistical tests for more than 2 samples

Jerry W. McLarty, PhD,* and Runhua Shi, MD, PhD*

INTRODUCTION

In a previous article statistical methods for comparing 2 groups were discussed. However, what if there are more than 2 groups? For example, how do you test for differences in the effectiveness of 3 treatments, say, drug A, drug B, and drug C? The 2-group t test and its equivalent nonparametric tests are tailored for comparing only 2 means or medians. A simple alternative comes to mind: compare drug A to drug B, drug A to drug C, and drug B to drug C. So, instead of 2 comparisons there are now 3. If there were 4 groups to test, it would require 6 two-group comparisons. For 5 groups, 10 comparisons would be required. Multiple t tests are not the preferred method of addressing the multiple-group comparison problem. The statistical problem with multiple t tests in this manner is discussed in more detail herein, and more powerful methods are introduced. Also, extension of χ^2 and other nonparametric tests to compare multiple groups simultaneously is described.

MULTIPLE COMPARISONS PROBLEM

The fundamental problem with using multiple t tests to compare more than 2 groups simultaneously is that the type I error rate (ie, probability of erroneously rejecting the null hypothesis by chance) is increased with each comparison. This is called the *multiple comparisons problem*. The problem can occur in a number of statistical testing situations. To illustrate how this happens, consider a study comparing the effectiveness of 3 different drugs. The null hypothesis is that the 3 drugs have an equal effect: $H_0: \mu_A = \mu_B = \mu_C$, where the μ 's are the respective means for each drug. If the probability of type I error $\alpha = .05$ for a single t test ($H_0: \mu_A = \mu_B$ or $H_0: \mu_B = \mu_C$ or $H_0: \mu_A = \mu_C$), the actual type I error rate for all 3 t tests combined is approximately 0.14. This means that 14 times of a hundred we would expect to reject the null hypothesis by chance, not 5 as for $\alpha = .05$. The overall type I error rate continues to increase with more tests: for 4 t tests it would be approximately 0.19. It can be seen that the type I error rate can become unacceptably large, and we would have less faith in our conclusion to reject the null hypothesis and claim that one of the drugs is best.

One simple method to get around the multiple comparisons problem is to adjust the P value: if there are n tests, reject the null hypothesis only if $P/n < .05$. In our 3-group example, a P value of $.05/3 \cong .017$ or less would be called significant.

This is called the Bonferroni correction, and it is commonly used. Gene array testing, for example, often performs tens of thousands of tests looking for polymorphisms linked to specific diseases. The Bonferroni corrected P values can be quite small in such testing (eg, 10^{-9} or lower).

The Bonferroni correction is simple to apply and is widely used, but it has been criticized as being too conservative, that is, more often than necessary it fails to reject the null hypothesis. Numerous other methods of adjusting for multiple comparisons have been proposed.

COMPARISON OF MEANS: ANALYSIS OF VARIANCE

The t test is a comparison of means test. As discussed in an earlier article, the t test can be used to compare one mean value (from a random sample) with a single number or to compare the mean from one group with the mean of another group or to compare 2 closely paired groups. The basic premise is that the difference in 2 means divided by the SE of the difference follows a well-known distribution, the t distribution. The t statistic can be seen as a signal to noise ratio (ie, the test statistic):

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{SE(\bar{X}_1)^2 + SE(\bar{X}_2)^2}} = \frac{\text{signal}}{\text{noise}}$$

which consists of the signal, the difference of the 2 means, and the noise, all that stuff in the denominator. This is a useful way of thinking about statistical tests: as the signal gets bigger (more difference between means), the test statistic gets larger and the P value is more significant. Conversely, if the noise in the system (ie, the variance or standard deviation (SD) of the samples) is large, the value of the test statistic decreases and the P value is less significant.

This same signal to noise analogy is the basis of a powerful statistical test, the analysis of variance (ANOVA). ANOVA in its simplest form is an extension of the t test to more than 2 groups. For ANOVA, the test statistic is called F and has its own probability distribution. Like the t test statistic, F is the ratio of 2 quantities:

F = mean square between groups/mean square within groups.

The derivation of this formula is given herein but may be skipped by readers more interested in an overview of ANOVA than computational details.

DERIVATION OF THE F TEST FORMULA

Mean Square Between Groups (Signal)

For the t test the difference in the 2 group means is the numerator of the formula; in ANOVA we have several group

Affiliations: * Department of Medicine, Feist-Weiller Cancer Center, Louisiana State University Health Sciences Center, Shreveport, Louisiana.

Disclosures: Authors have nothing to disclose.

Received for publication February 16, 2009; Received in revised form April 2, 2009; Accepted for publication April 3, 2009.

means to compare, so the trick is to calculate the difference between each group mean and the overall mean (the mean of all the values in all groups). If the overall mean is $\bar{X}_{ALL}$, and $\bar{X}_A$, $\bar{X}_B$, and $\bar{X}_C$ are 3 group means we wish to compare, then $SS_{\text{Between Groups}} = n[(\bar{X}_A - \bar{X}_{ALL})^2 + (\bar{X}_B - \bar{X}_{ALL})^2 + (\bar{X}_C - \bar{X}_{ALL})^2]$, where n is the number of subjects in each group and SS is the abbreviation for sum of squares. If we added just the 3 differences without first squaring them, they would cancel out and sum to 0, which would give us no information about the signal. To avoid this, the differences are squared, hence the term *sum of squares*.

Only one new concept needs to be introduced, that is, mean square. If we wanted to get an average squared difference it would be intuitive to divide the SS described previously by 3 (3 group differences were obtained). However, as with SD , to get an unbiased estimate using our sampled groups, we divide the sum of squares by the degrees of freedom. One degree of freedom was used in calculating the overall mean. So, to get our average, we divide the sum of squares by the number of groups minus 1, that is, $k - 1$; in this example, $df = 3 - 1 = 2$. In general, the between-groups degrees of freedom ($df_{\text{Between Groups}}$) = $k - 1$, where k is the number of groups. Finally, we calculate the average, called the mean square (abbreviated as MS):

$$MS_{\text{Between Groups}} = SS_{\text{Between Groups}} / df_{\text{Between Groups}}$$

This is the numerator (signal) portion of our F statistic.

Mean Square Within Groups (Noise)

The goal of ANOVA is to compare means between groups. There is variability between groups and variability within each group. It is this variability within groups that leads to the denominator (noise) portion of our F statistics.

For our example with 3 groups, let's assume that the sample size is 10 individuals for each group or 30 people total. If we compute the squared difference between every data point and its group mean and then add them, we would have a sum of squares within groups:

$$SS_{\text{Within Groups}} =$$

$$(X_{1A} - \bar{X}_A)^2 + (X_{2A} - \bar{X}_A)^2 + \dots + (X_{10A} - \bar{X}_A)^2 + (X_{1B} - \bar{X}_B)^2 + (\bar{X}_{2B} - \bar{X}_B)^2 + \dots + (X_{1C} - \bar{X}_C)^2 + \dots + (\bar{X}_{2C} - \bar{X}_C)^2 + \Lambda + (\bar{X}_{10C} - \bar{X}_C)^2$$

where X_{1A} , for example, means the value of the first data point in group A, and $\dots$ denotes terms omitted for brevity.

For our example, there was a total sample size of 30. Three degrees of freedom were used to calculate the 3 group means. So the overall within-group degrees of freedom is as follows: $df_{\text{Within Groups}} = N - k$ (where N is the total sample size for all groups combined and k is the number of groups). For our example, $df_{\text{Within Groups}} = 30 - 3 = 27$. In a manner similar to between groups, the within-groups mean square is as follows:

$$MS_{\text{Within Groups}} = SS_{\text{Within Groups}} / df_{\text{Within Groups}}$$

Calculation of F

Now that both our numerator (signal) and denominator (noise) have been defined, we can calculate the test statistic as follows:

$$F = MS_{\text{Between Groups}} / MS_{\text{Within Groups}}$$

With the test statistic F , the P value can be calculated with a computer or looked up in a table with numerator degrees of freedom $k - 1$ and denominator degrees of freedom $N - k$.

Example (Hypothetical)

Assume that diastolic blood pressure is measured in 5 subjects in each of 3 age groups: 30 to 39, 40 to 49, and 50 to 59 years. The question is, "Does diastolic blood pressure differ in these age groups?" The null hypothesis is the group means are equal. The data table looks like this:

Subject No.	Age 30-39 years	Age 40-49 years	Age 50-59 years
1	70	80	90
2	75	70	85
3	60	75	80
4	70	85	95
5	80	80	80
Group mean	71.0	78.0	86.0

The overall mean is 78.3 mm Hg.

The sum of squares between groups is as follows:

$$SS_{\text{Between Groups}} = 5[(71-78.3)^2 + (78-78.3)^2 + (86-78.3)^2] = 563.3$$

$$MS_{\text{Between Groups}} = SS_{\text{Between Groups}} / df_{\text{Between Groups}} = 563.3 / 2 = 281.7$$

$$SS_{\text{Within Groups}} = (70-71)^2 + (75-71)^2 + \dots + (80-78)^2 + (70-78)^2 + \dots + (90-86)^2 + \dots + (80-86)^2 = 520.0$$

$$MS_{\text{Within Groups}} = SS_{\text{Within Groups}} / df_{\text{Within Groups}} = 520 / 12 = 43.3$$

Finally,

$$F = MS_{\text{Between Groups}} / MS_{\text{Within Groups}} = 281.7 / 43.3 = 6.5$$

The P value for $F = 6.5$ with 2 and 12 $df = 0.012$.

Because P is less than .05, we reject the hypothesis that the group means of diastolic blood pressure are equal.

Which Groups Are Different?

In this example, the F test tells us that the groups means are not all the same (ie, at least one of them is different from another). Unfortunately, it does not tell us which one(s) is (are) different. In the example, it is easy to see that the youngest age group has a much lower mean diastolic blood pressure than does the oldest, but it is not clear whether the middle age group significantly differs from the other two. To address this issue, there are several post hoc tests (tests to be used after the null hypothesis has been rejected) that can help sort this out. Post hoc tests are beyond the scope of this

discussion, but most statistical software has several of them included with the ANOVA procedure.

Comparison of ANOVA and the t Test

The underlying assumptions of the *t* test apply to ANOVA: the groups are assumed to be independent random samples from normal populations. In its simplest form, ANOVA is just an extension of the *t* test to more than 2 groups. If ANOVA is performed with just 2 groups, the *P* value will be identical to the *P* value of a *t* test on the same data.

Extensions of Simple ANOVA

ANOVA is a powerful statistical tool with capabilities far beyond expanding *t* test methods to more than 2 groups. It is possible to have more than one factor (in the example given age group is a factor) in ANOVA. For example, a similar study might include sex and race and even include interaction effects among them; it might be that drugs work differently for older African American men than for white men and women and African American women. In this case there would be a significant age-race-sex interaction effect. Factors in ANOVA are variables of interest that can be categorized into a few values, such as age group, sex, race, and study center. A single ANOVA test can test all of the factors at once. The basic principals of comparing between and within mean squares are the same as for the simple 1-factor ANOVA.

Analysis of covariance (ANCOVA) extends this concept to continuous variables (called *covariates*), such as age, height, and cholesterol level. Again, ANCOVA tests all factors and covariates and interactions in a single run. Generally, covariates are tested first, and if significant, the factors are tested after adjusting for the covariates. For example, if age is a covariate and sex is a factor, the effect of sex may be computed after adjusting for age, simulating the effect that all subjects are the same age.

Another powerful extension of ANOVA is repeated-measures ANOVA. In many studies, the same outcome is measured multiple times on the same individuals. For example, the effect of an allergy treatment might require measuring forced expiratory volume in 1 second (FEV₁) at weekly intervals for 5 weeks. The investigators could just compare the effect of the drug, compared with a placebo, by using the FEV₁ at week 5. However, this ignores information that may be useful from the first 4 weeks of data. Repeated-measures ANOVA is a means of using all the data from all of the time points.

It is possible that the sample sizes are not always equal for each group. However, the same methods described herein still apply, but the computation of degrees of freedom within groups is different. Most common software packages take care of this automatically. In general, ANOVA calculations are sufficiently tedious so it is not practical to do them by hand when there are so many software packages readily available. Statistical software will be discussed in a later

article. Readers interested in more advanced applications of ANOVA should consult statistical textbooks.

COMPARING PROPORTIONS IN MORE THAN 2 GROUPS

Two-sample comparisons of proportions have been described in a previous article. Expansion of the χ^2 test to more than 2 samples (groups) involves no new concepts. The principle of comparing observed to expected values in each cell of a table is identical regardless of the number of samples, and the rules of thumb concerning the magnitude of expected values are the same.

Recall that for each cell in a table the comparison of the observed data to the expected data (based on the assumptions that rows and columns are independent) follows this format:

$$(\text{Observed} - \text{Expected})^2 / \text{Expected}$$

The differences are squared to eliminate the sign of the difference. Otherwise the sum of differences used in the following equation would always sum to zero (ie, pluses and minuses would cancel each other out). The χ^2 test statistic is the sum of all such squared terms in the table. In mathematical symbols this becomes the following:

$$\chi^2 = \sum_i \sum_j \frac{(\text{observed value} - \text{expected value})^2}{\text{expected value}} = \frac{\sum_i \sum_j (O_{ij} - E_{ij})^2}{E_{ij}}$$

where *i* = 1,2,3, ..., *N*_{rows} indexes rows and *j* = 1,2,3,...*N*_{cols} indexes columns.

If appropriate conditions are met, then this sum approximates a χ^2 distribution and probabilities (*P* values) can be calculated. The degrees of freedom for calculating *P* values is (*N*_{rows} - 1) × (*N*_{cols} - 1). There is no restriction on how large the number of rows and columns can be, except there must be at least 2 cells in the table. As before, no expected values can be less than 1.0, and 20% or more of the expected values must be 5.0 or greater. If these rules are violated, the approximation to the χ^2 distribution is suspect and the conclusions reached may be in error. This can often be fixed by collapsing rows or columns (ie, combining 2 or more groups into 1 group). If it is not possible to do this, there is a test called the Fisher exact test that is not restricted by small or even zero expected values. However, this can be computationally intensive, depending on the size of the table and the number of subjects. To illustrate the χ^2 test, consider the following example, an extension of the example used in an earlier section.

Assume we want to compare the gene frequencies of a gene mutation possibly involved in an allergic disease for 3 different racial groups: Asian, non-Hispanic white, and Hispanic white. The data are as follows:

Subjects	AA	AG	GG	Total No. of subjects
Asian	15 (12.2)	60 (47.1)	30 (45.7)	105
Non-Hispanic white	10 (13.8)	44 (53.3)	65 (51.8)	119
Hispanic white	3 (2.0)	4 (7.6)	10 (7.4)	17
Total	28	108	105	241

The numbers in parentheses are expected values. For tables like this, the expected values for a given cell (say the AA genotype in Asians) is column total (28) times row total (105) divided by the overall total (241). This assumes that the column values are distributed proportionately in each row.

The χ^2 for this table is 18.8, with 4 *df*. All of the expected values are greater than 1.0, and only one cell had an expected value below 5.0. The χ^2 assumptions are valid for this table. The *P* value is .001. We conclude that there are racial differences in the distribution of genotypes, with Asians having a predominance of the AG genotype.

NONPARAMETRIC COMPARISONS OF MORE THAN 2 GROUPS

As we have seen, ANOVA assumes that the dependent (outcome) variable is drawn from a normal distribution. When this assumption is not met, nonparametric methods based on ranks have been developed. The 2-sample case has been described in a previous article. The 2-sample case is called the Mann-Whitney test. An equivalent test for more than 2 groups is the Kruskal-Wallis test. The principal idea behind the Kruskal-Wallis test is to rank all the data without regard to group and then calculate an overall average rank and an average rank within each group. Then the difference between group mean ranks and the overall mean rank is calculated and summed over all groups, much like the sum of squares for ANOVA. The formula for the test statistic *K* is as follows:

$$K = \frac{12}{N(N+1)} \sum_{i=1}^g n_i (\bar{r}_i - \bar{r})^2$$

where *N* is the total number of subjects, *g* is the number of groups, *n_i* is the number of subjects in group *i*, $\bar{r}_i$ is the mean rank for group *i*, and $\bar{r}$ is the overall mean rank.

For group sizes of 5 or more, *K* has an approximate χ^2 distribution with *g* - 1 *df*. To illustrate, let's apply the Kruskal-Wallis test to the following example. Assume we want to compare the prostate specific antigen (PSA) levels in 3 groups of men ages 40 to 50, 50 to 60, and 70 to 80 years. This is a hypothetical example, but PSA is a good illustration: it is not normally distributed but is highly skewed.

Raw Data

Subject No.	Age 40-49 years	Age 50-59 years	Age 60-69 years
1	0	0.4	1.2
2	0.1	0.6	2.0
3	0.2	0.8	3.0
4	0.3	1.0	2.4
5	0.5	1.4	4.6

Ranked Data

Subject No.	Age 40-49 years	Age 50-59 years	Age 60-69 years
1	1	5	10
2	2	7	12
3	3	8	14
4	4	9	13
5	6	11	15

The overall mean rank is 8.0. There are 3 groups and a total of 15 subjects. So,

$$K = \frac{12}{15(16)} [5(3.2 - 8)^2 + 5(8 - 8)^2 + 5(12.8 - 8)^2]$$

K = 11.52, *df* = 3 - 1 = 2, *P* = .003.

So we reject the null hypothesis of no age difference in PSA.

The problem of which groups are different remains. Unfortunately, the methods are more limited than for ANOVA. A simple alternative, after determining from the Kruskal-Wallis test that the groups are not the same, is comparing pairs of groups using the Mann-Whitney test and reducing the type I error rate from 0.05 to 0.05 divided by the number of comparisons (Bonferroni correction). In this example, we would reject the pairwise null hypothesis with *P* < .05/3 or .0167.

REFERENCES

Included in a comprehensive list at the end of the Supplement.

Requests for reprints should be addressed to:

Jerry W. McLarty, PhD

Department of Medicine

Feist-Weiller Cancer Center

LSU Health Sciences Center

1501 Kings Highway

Shreveport, LA 71103-4228

E-mail: JMCLAR@LSUHSC.EDU