CHAPTER SEVEN

MULTIPLE COMPARISONS

As we saw in the last chapter, a common statistical method for comparing the means from two groups of patients is the t-test. Frequently, however, we wish to compare more than two groups of patients in order to determine whether a difference exists between any or all of the groups involved. There are several statistical methods for simultaneously comparing several groups of patients, all of which are examples of **multiple comparison procedures**.

MULTIPLE T-TESTS

One option for comparing three or more groups of patients is to perform two-sample, unpaired t-tests on each of the possible pairwise combinations of the data, and compare the resulting p-values. As we will see, however, this method is not appropriate as the t-test is designed to evaluate differences between only two groups of patients. The use of multiple t-tests in this manner is one of the most commonly seen statistical errors in the medical literature.

For example, consider a study in which we randomly assign 100 patients to receive one of four different antibiotics (Drug A, B, C, and D) prior to operation, and we wish to assess the efficacy of each drug in preventing post-operative wound infections. In order to analyze the four groups of data using t-tests, we would need to perform two-sample, unpaired t-tests on each of the following 6 pairwise comparisons:

Drug A vs Drug B	Drug A vs Drug C	Drug A vs Drug D
Drug B vs Drug C	Drug B vs Drug D	Drug C vs Drug D

The problem with this approach is that with each comparison we make, there is always a chance, however small, that we will commit a Type I error; that is, we will erroneously reject the null hypothesis when, in reality, there is no difference between the groups. In a single two-sample t-test with a significance level of 0.05, the likelihood of making such a Type I error is only 5%. This is known as the **per-comparison error rate**. However, if we use the same two-sample t-test to evaluate all four groups of data and perform all 6 possible pairwise comparisons, the likelihood of making a Type I error in at least one of our 6 comparisons rises to 30% (0.05 x 6). This is known as the **per-experiment error rate**. Instead of having 95% confidence that our conclusions are correct, our confidence is now decreased to 70% and we are more likely to commit a Type I error. Thus, the two-sample t-test can lead to erroneous conclusions if we improperly use it to make multiple comparisons.

We could still use t-tests to perform multiple comparisons, acknowledging the increased per-experiment error rate, if it weren't for another problem. The use of multiple t-tests results in the calculation of multiple p-values (one for each comparison) which can only be used to compare the two groups within each comparison. There is not a separate p-value which we can use to compare all of the groups simultaneously and thereby document that one therapy or treatment is better than the rest. Thus, using t-tests we still cannot compare more than two groups of patients at a time. Multiple t-tests should therefore not be used in the statistical analysis of more than two groups of data.

BONFERRONI ADJUSTMENT

As noted above, if we use two-sample statistical tests to perform multiple comparisons, the potential for error (the per-experiment error rate) is additive such that we are more likely to make erroneous conclusions. The **Bonferroni adjustment** takes this increase in the per-experiment error rate into account by adjusting the per-comparison error rate downward so that the likelihood of making a Type I error with each comparison is decreased. The Bonferroni adjustment allows us to ask the question "Are all of the groups different from each other?"

For example, in our study on antibiotic efficacy we were interested in comparing four groups of patients. If we wished to determine that all four groups were different from one another with 95% confidence, our overall

per-experiment error rate would need to be 0.05. To determine the per-comparison error rate for each test using the Bonferroni adjustment, we would divide our desired per-experiment error rate by the number of comparisons. For this example, the per-comparison error rate for each t-test would then be 0.05/6 or 0.0083. We would thus perform the six t-tests using a significance level of 0.0083. If all six t-tests resulted in a p-value of less than 0.0083, we could then state that all four groups were statistically different from one another with 95% confidence (a significance level of 0.05).

One problem with the Bonferroni method is that it only estimates the true per-experiment error rate. The actual chance of making a Type I error may be much less. Consider the case where one of our four antibiotics (Drug B) is much more effective in preventing wound infections than are the other three (whose efficacies are all very similar). If we perform two-sample t-tests on each of the six possible combinations, we might obtain the following results (note that our analysis results in six different p-values which cannot be used to evaluate the study as a whole):

Wound Infection Rates for Drugs A, B, C, and D

Comparison	p-value	Comparison	p-value
Drug A vs Drug B	0.01	Drug B vs Drug C	0.02
Drug A vs Drug C	0.20	Drug B vs Drug D	0.04
Drug A vs Drug D	0.50	Drug C vs Drug D	0.60

Using the Bonferroni adjustment, in order to have 95% confidence in our results, our per-comparison significance level must be 0.0083 (0.05/6) and our hypotheses would be as follows:

Null Hypothesis: none of the drugs prevent wound infections
Alternate Hypothesis: all four drugs prevent wound infections

Based on the Bonferroni adjustment, in order to reject our null hypothesis with 95% confidence, each of the six p-values must be less than 0.0083. Since this is not the case, we must accept our null hypothesis and conclude that none of the drugs are efficacious in preventing wound infections. This is clearly not the case, however, as the efficacy of Drug B is significantly greater than that of Drugs A, C, and D. In this situation, use of the Bonferroni adjustment results in our ignoring the significant differences present. The Bonferroni adjustment, by being a very conservative statistical test, loses statistical power and is more likely to result in a Type II error. By lowering the per-comparison error rate, it reduces the likelihood of erroneously concluding that a difference exists for the experiment as a whole (a Type I error), but at the same time makes it more likely that a significant difference among the groups will be missed (a Type II error).

ANALYSIS OF VARIANCE (ANOVA)

A common solution to the issue of comparing three or more groups is a test known as **analysis of variance** or **ANOVA**. It addresses the question of whether there are differences between the means of the groups. It does not, however, identify which of the groups differ from one another. It is a method which expands on the traditional t-test allowing evaluation of multiple groups of observations without the increased risk of a Type I error.

Like the t-test, ANOVA makes three assumptions. First, the observations are assumed to be normally distributed. If this is not the case, the data must first be transformed to a normal distribution or a non-parametric multiple comparisons method must be utilized. Second, the population variance is assumed to be the same in each group. The importance of this assumption is lessened if the sample sizes are equal. Third, the observations in each group must be independent and cannot affect the values of observations in another group. As with any statistical analysis, the raw data should be examined initially to determine whether these assumptions are met. ANOVA is a method that is complex, but is readily performed on a personal computer via a statistics package or spreadsheet. In order to understand the various types of ANOVA and the results they provide, we will briefly discuss the theory behind performing such an analysis.

ANOVA begins by calculating the mean of each group of data. It then combines the data from each group into a single group and calculates the **grand mean** of the grouped data. ANOVA uses these means to ask two questions:

1) Is there a difference between the groups (the **between-groups variance**)?

- If the group means are similar to the grand mean of all of the data, the variance of the observations within the groups will be small and the groups will likely be very similar.

2) How much variability exists between the observations and their respective group means (the **within-groups variance**)?

- If the variability between the groups is similar to the variability within the groups, they are likely from the same population.

An **F test** is then performed, resulting in a critical value similar to that obtained for a t-test. The null hypothesis of the F test is that the group means are not different. If the ratio of the between-groups variance to within-groups variance is small, the group means will likely be very similar. The calculated F test statistic will then be small and will not reach significance. We would then accept the null hypothesis concluding that the groups are not different. If the ratio of the variances is large, the means of the groups are likely to be different. The calculated F statistic will be large and will likely exceed the critical value of F necessary to determine a significant difference. We would then accept the alternate hypothesis and state that a difference exists.

It is important to remember that a significant F test only indicates that some or all of the group means are different. As we do not know which of the means are different, further analyses known as **post hoc comparisons** must be performed to determine how the groups differ from one another. Examples of such methods include Tukey's HSD procedure, Scheffe's procedure, the Newman-Keuls procedure, and Dunnett's procedure, all of which are easily performed by a computer statistics package.

A computer generated ANOVA output usually includes the **sum of squares**, the **mean square**, the **degrees of freedom (df)** for the **between-groups and within-groups variance** calculations, the **F test statistic**, and the **significance of the F test**. The sum of squares refers to the sum of the squared differences between the group means. If the means are similar, the sum of squares will be small (and vice versa). The mean square for the between-groups variance is the sum of squares divided by the degrees of freedom, and is a measure of the variation of the between-groups means around the grand mean. The mean square for the within-groups variance (also known as the **error mean square**) is a pooled estimate of the variation of the within-groups observations around their respective group means. The F statistic is calculated by dividing the mean square of the between-groups variance by the mean square of the within-groups variance. Critical values of the F statistic can be obtained from an F distribution table based on the number of groups being compared (k) and the number of observations present in each group (n). The degrees of freedom for the denominator (the between-groups variance) is given by k - 1. The degrees of freedom for the numerator (the within-groups variance) is given by k(n - 1).

ANOVA exists in two forms. **One-way or single-factor ANOVA** compares two or more independent variables with a single dependent or outcome variable. An example would be our study comparing the efficacy of four antibiotics (the independent variables) on the incidence of wound infection (the dependent variable). **Two-way or two-factor ANOVA** allows for evaluation of the effect of two or more independent variables on two different dependent variables. We would use two-factor ANOVA if we wished to determine not only the efficacy of the four antibiotics on the incidence of wound infections (the first dependent variable), but also their impact on hospital length of stay as a result of wound infections (the second dependent variable).

NON-PARAMETRIC ANOVA

One of the primary assumptions in the use of ANOVA is that the data are normally distributed. If this is not the case, the use of ANOVA may result in erroneous conclusions. When data are not normally distributed, there are two options. The first option is to transform the data to a normal distribution using a logarithmic or square root transformation. One problem with this method is that the resulting ANOVA will refer to the transformed observations and not the original data. Further, the units of transformed data may be difficult to interpret due to the logarithmic manipulation of the data. The second option for dealing with non-normally distributed data is to perform a **non-parametric ANOVA**. Like the non-parametric versions of the t-test (the Wilcoxon signed-ranks test and Wilcoxon rank-sum test), non-parametric ANOVA is based on analysis of the ranks of the data. The non-parametric equivalent of one-way ANOVA is the **Kruskal-Wallis test** while the equivalent of two-way ANOVA is the **Friedman two-way ANOVA by ranks**. These tests can be found in most computer statistics packages.

META-ANALYSIS

A special case of multiple comparisons is **meta-analysis**. Meta-analysis is a statistical process for systematically reviewing and combining the results of multiple studies in order to draw conclusions about a treatment or outcome based on the combined data and increased sample size. It is helpful when 1) the existing studies are in disagreement as to a conclusion, 2) the existing studies have small sample sizes and therefore lack statistical power, 3) the effect size of interest is small and requires a larger number of observations to reach statistical significance, or 4) a large scale trial would prove too costly to perform. The conclusions of meta-analyses are frequently helpful in planning future large scale clinical trials.

Meta-analyses are being seen more and more commonly in the medical literature. They must be carefully designed and the outcome of interest clearly formulated before the study is begun in order to ensure accurate conclusions. Meta-analysis begins by identifying all relevant clinical trials and studies pertaining to the subject of interest not only through computerized literature searches (which may contain only 60% of relevant studies), but also through review of textbooks, referenced articles, published abstracts, and unpublished research. A major concern in meta-analysis is **publication bias**. “Negative studies” (i.e., those that do not have a positive finding) are less likely to be published than “positive studies”. These unpublished results, had they been published, could potentially alter the results of a meta-analysis thereby ntroducing bias. Statistical methods exist to take into account the potential effect of unpublished negative studies on meta-analysis results.

Clinical trials and studies which are identified from the medical literature are then carefully assessed to determine their “combinability” (i.e., are the trials sufficiently similar that they can be compared?) and “homogeneity” (i.e., are the patients evaluated in each study similar). Careful definition of inclusion and exclusion criteria is essential in identifying those studies which will be analyzed. The data quality and design of each study is also assessed. Missing or insufficient data may require that the original authors be contacted to obtain the necessary information to ensure accurate statistical analysis and clinically useful conclusions.

The data are subsequently blinded and reviewed critically by at least two investigators to confirm that the data are being interpreted accurately. Special statistical methods are utilized to “pool” the data from each trial and determine whether the proposed study hypotheses can be either accepted or rejected with significance. **Sensitivity analysis** is then performed to ascertain whether inclusion (or exclusion) of various types of studies (randomized versus non-randomized, controlled versus uncontrolled) makes a difference in the meta-analysis conclusions. Meta-analysis, although methodologically complex and time consuming, is a useful statistical method for reaching conclusions that would otherwise not be possible in a single clinical trial. It will likely be seen more and more frequently in the medical literature as an economic alternative to performing expensive, multi-institutional studies. In order for such meta-analyses to be valid, however, strict, rigorous guidelines must be followed to ensure that the meta-analysis is performed correctly. The reader is referred to the references at the end of this chapter for further details on the methodology and interpretation of meta-analyses.

SUGGESTED READING

1. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 1. introduction. Mayo Clin Proc 1988;63:813-815.
2. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 2. comparisons among several therapies. Mayo Clin Proc 1988;63:816-820.
3. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 3. repeated measures over time. Mayo Clin Proc 1988;63:918-920.
4. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 4. performing multiple statistical tests on the same data. Mayo Clin Proc 1988;63:1043-1045.
5. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 5. comparing two therapies with respect to several endpoints. Mayo Clin Proc 1988;63:1140-1143.
6. Godfrey K. Comparing the means of several groups. NEJM 1985;313:1450-1456.
7. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials: a survey of three medical journals. NEJM 1987;317:426-432.
8. Dawson-Saunders B, Trapp RG. Basic and clinical biostatistics (2nd Ed). Norwalk, CT; Appleton and Lange, 1994:125-141.
9. Wassertheil-Smoller S. Biostatistics and epidemiology: a primer for health professionals. New York; Springer-Verlag, 1992:46-53.
10. L'Abbe KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. Ann Intern Med 1987;107:224-233.
11. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. NEJM 1987;316:450-455.
12. Thacker SB. Meta-analysis: a quantitative approach to research integration. JAMA 1988;259:1685-1689.
13. Detsky AS, Baker JP, O'Rourke K, Goel V. Perioperative parenteral nutrition: a meta-analysis. Ann Intern Med 1987; 107:195-203.
14. Chalmers TC, Smith H, Blackburn B, Silverman B, Schroeder B, et al. A method for assessing the quality of a randomized control trial. Controlled Clin Trials 1981; 2:31-49.
15. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. JCNI 1959; 22:719-748.
16. Rosenthal R. The "file drawer problem" and tolerance for null results. Psychol Bull 1979; 86: 638-641.
17. Begg CB. A measure to aid in the interpretation of published clinical trials. Stats Med 1985; 4:1-9.
18. Easterbnrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet 1991; 337:867-872.