0.1 Introduction

An ANOVA F-test will reveal whether the means, \( \mu_i \), are different. However, the experimenter usually wants to know which \( \mu_i \) are significantly different from which others, multiple comparison procedures are rules which provide such information.

In the following, five different multiple comparison procedures will be studied; Tukey’s studentized range, Scheffe’s multiple comparison procedure, Bonferroni t-intervals, and Dunnett’s test, all of which result in simultaneous confidence intervals to test for significant differences. Other tests, called multiple range tests which are more powerful than those previously mentioned, include Duncan’s multiple range and the Student- Newman-Keuls procedure. For the one-way analysis of variance model, with fixed effects, each method’s computational formula is given and then compared to the other similar methods based on applicability, error rates, and conservativeness.

0.2 Terminology

A procedure is called conservative if the actual simultaneous confidence level is greater than the stated confidence level.

A Type I error occurs when a procedure determines two means to be significantly different when in fact, they are the same. A Type II error occurs when a procedure determines two means the same, when they are actually different. The power is defined to be equal to \( 1 - P(\text{Type II error}) \).

The comparisonwise error rate is "the probability of making a Type I error for a specific comparison". Procedures which control for the comparisonwise error rate tend to be more liberal (non-conservative), in that they will reject the hypothesis that two means are the same with less evidence. However, this also implies that differences in two means will be detected more easily, and so procedures which control for comparisonwise errors will be more powerful.

The experimentwise error rate is "the probability of making at least one Type I error". Procedures which control the experimentwise error rate will tend to be conservative, since more evidence is needed to reject the hypothesis that two means are the same.
0.3 Tukey’s Studentized Range

Tukey’s method is used when the set of all pairwise comparisons of means is of interest. This method is constructed using the studentized range distribution. This distribution can be derived by taking $Y_1, \ldots, Y_r$ to be $r$ independent observations from a Normal($\mu, \sigma^2$) distribution. Let

$$w = \max(Y_i) - \min(Y_i),$$

and let $s^2$ be the estimate of $\sigma^2$ that is independent of the $Y_i$ and based on $\nu$ degrees of freedom. Then $\frac{w}{s}$ is the studentized range statistic and is denoted by,\(^1\)

$$q(r, \nu) = \frac{w}{s}.$$

Consider the ANOVA model

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where $i = 1, \ldots, r$ and $j = 1, \ldots, n_i$, $N = \sum_i n_i$, $\varepsilon_{ij}$ are normal(0,$\sigma^2$), and $s^2$ is the estimate of $\sigma^2$. The Tukey multiple comparison confidence intervals for the $\frac{r(r-1)}{2}$ pairwise comparisons $\mu_i - \mu_j$, $i \neq j$ are then given by

$$\bar{Y}_i - \bar{Y}_j \pm q(\alpha; r, N-r) s \left( \frac{1}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right) \right)^{1/2}$$

where $q(\alpha; r, N-r)$ is the upper $\alpha$ point of the Studentized range distribution with $r$ and $N - r$ degrees of freedom. The above reduces to the correct form for equal sample sizes by taking $n = n_i$ for all $i$.

And so, $\mu_i$ and $\mu_j$ are considered to be significantly different if the Tukey confidence interval for $\mu_i - \mu_j$ does not contain zero. Therefore, the Tukey procedure will make a Type I error only if zero is not in some Tukey confidence interval when $\mu_i - \mu_j = 0$.\(^2\)

When all sample sizes are equal, the experimentwise error rate is controlled, that is

$$P(\mu_i - \mu_j \in (\bar{Y}_i - \bar{Y}_j \pm q(\alpha; r, N-r) s \left( \frac{1}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right) \right)^{1/2}) \forall i \neq j) = 1 - \alpha.$$

However, when sample sizes are not equal than

$$P(\mu_i - \mu_j \in (\bar{Y}_i - \bar{Y}_j \pm q(\alpha; r, N-r) s \left( \frac{1}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right) \right)^{1/2}) \forall i \neq j) > 1 - \alpha,$$

\(^1\)Neter, Wasserman, and Kutner, Applied Linear Statistical Models, p. 381
\(^2\)Arnold, The Theory of Linear Models and Multivariate Analysis, p. 187
in other words, the Tukey method is conservative.\textsuperscript{3}

In addition,

If not all pairwise comparisons are of interest, the confidence coefficient for the family of comparisons being considered will be greater than the specification $1 - \alpha$ used in setting up the Tukey intervals. Thus, the confidence coefficient $1 - \alpha$ with the Tukey method serves as a guaranteed minimum level when not all pairwise comparisons are of interest.\textsuperscript{4}

\textsuperscript{3}Neter, Wasserman, and Kutner, \textit{Applied Linear Statistical Models}, p. 580

\textsuperscript{4}Neter, Wasserman, and Kutner, \textit{Applied Linear Statistical Models}, p. 584
0.4 Scheffe's Multiple Comparison Procedure

The Scheffe method is used when we wish to estimate the set of all possible contrasts

\[ L = \sum_{i=1}^{r} c_i \mu_i \]

of \( \mu_1, \ldots, \mu_r \) where \( \sum_{i=1}^{r} c_i = 0 \). Note that infinitely many statements belong to this family and that the pairwise comparisons for \( \mu_j - \mu_k \) is found by taking \( c_j = 1, c_k = -1, \text{and } c_i = 0 \) for all \( i \neq j \) and \( i \neq k \).

Consider the ANOVA model

\[ Y_{ij} = \mu_i + \epsilon_{ij} \]

where \( i = 1, \ldots, r \) and \( j = 1, \ldots, n_i \), \( N = \sum_{i=1}^{r} n_i \), \( \epsilon_{ij} \) are \( \text{normal}(0, \sigma^2) \), and \( s^2 \) is the estimate of \( \sigma^2 \). The Scheffe method joint confidence interval for the contrast \( L = \sum_{i=1}^{r} c_i \mu_i \) is given by

\[ \sum_{i=1}^{r} c_i \bar{Y}_i \pm s((r-1)F(\alpha, r-1, N-r)\sum_{i=1}^{r} \frac{c_i^2}{n_i})^{1/2} \]

where \( F(\alpha, r-1, N-r) \) is the upper \( \alpha \) point of the central \( F \) distribution with \( r-1 \) and \( N-r \) degrees of freedom.

In addition,

\[ P(\sum_{i=1}^{r} c_i \mu_i \in \left( \sum_{i=1}^{r} c_i \bar{Y}_i \pm s((r-1)F(\alpha, r-1, N-r)\sum_{i=1}^{r} \frac{c_i^2}{n_i})^{1/2} \right) \forall \text{contrasts}) = 1-\alpha \]

whether the sample sizes are equal or unequal.\(^5\) However,

Since applications of the Scheffe method never involve all conceivable contrasts, the confidence coefficient for the finite family of statements actually considered will be greater than \( 1-\alpha \).

Thus, when we state the confidence coefficient is \( 1-\alpha \) with the Scheffe method, we really mean it is guaranteed to be at least \( 1-\alpha \).\(^6\)

As with the Tukey procedure, \( \mu_i \) and \( \mu_j \), are considered significantly different if the Scheffe confidence interval for \( \mu_i - \mu_j \) does not contain zero.

---

\(^5\)Neter, Wasserman, and Kutner, Applied Linear Statistical Models, p. 585
\(^6\)Neter, Wasserman, and Kutner, Applied Linear Statistical Models, p. 587
Therefore, the Scheffe method makes a Type I error if zero is not in some Scheffe confidence interval when in fact, $\mu_i - \mu_j = 0$. Since the probability of all the simultaneous intervals being correct is $1 - \alpha$, then the probability of a Type I error is at most $\alpha$.\(^7\)

Therefore, Scheffe's test is very conservative, and so this test may not be as effective as desired if one is interested only in a small group of contrast, such as the set of pairwise comparisons between means.

\(^7\)Arnold, *The Theory of Linear Models and Multivariate Analysis*, p. 187
0.5 Bonferroni $t$-Intervals

The Bonferroni method is applicable when we wish to investigate a particular set of contrasts (i.e. the set of all pairwise comparisons).

Consider the ANOVA model

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where $i = 1, \ldots, r$ and $j = 1, \ldots, n_i$, $N = \sum_{i=1}^{r} n_i$, $\varepsilon_{ij}$ are normal$(0, \sigma^2)$, and $s^2$ is the estimate of $\sigma^2$. The Bonferroni confidence intervals for the $g$ contrasts $L^{(1)}, \ldots, L^{(g)}$, where $L^{(k)} = \sum_{i=1}^{r} c_i^{(k)} \mu_i$, are given by

$$\sum_{i=1}^{r} c_i^{(k)} \bar{Y}_i \pm t\left(\frac{\sigma}{2g}, N - r\right) s\left(\sum_{i=1}^{r} \frac{c_i^{(k)}}{n_i}\right)^{\frac{1}{2}}$$

where $t\left(\frac{\sigma}{2g}, N - r\right)$ is the upper $\alpha$ point of the central $t$-distribution with $N - r$ degrees of freedom.

The probability of all such intervals being simultaneous correct is at least $1 - \alpha$. This can be shown by use of the Bonferroni inequality.

Suppose $A_i$, $i = 1, \ldots, g$ is the event that the $i^{th}$ statement is correct, where $P(A_i) = 1 - \frac{\alpha_i}{g}$ and $P(A_i^c) = \frac{\alpha_i}{g}$. Then

$$P(\bigcap A_i) = 1 - P(\bigcap A_i^c) = 1 - P(\bigcup A_i^c)$$

$$\geq 1 - \sum_{i=1}^{g} \frac{\alpha_i}{g}$$

and taking $\alpha_i = \alpha$ yields

$$P(\bigcap A_i) = 1 - g\left(\frac{\alpha}{g}\right) = 1 - \alpha.$$ 

Therefore, the probability of all intervals being correct is at least $1 - \alpha$.

The above shows the superiority (when we are interested in controlling the experimentwise error rate) of the Bonferroni method over individual comparisons using the $t$-statistics based on the $t\left(\frac{\sigma}{2}, N - r\right)$ distribution. In this situation, taking $P(A_i) = 1 - \alpha_i$ and $P(A_i^c) = \alpha_i$, then the same method used above implies

$$P(\bigcap A_i) = 1 - P(\bigcap A_i^c) = 1 - P(\bigcup A_i^c)$$
\[ \geq 1 - \Sigma P(A_i^c) = 1 - \sum_{i=1}^{g} \alpha_i \]

and taking \( \alpha_i = \alpha \) yields

\[ P(\bigcup A_i) = 1 - g(\frac{\alpha}{g}) = 1 - \alpha. \]

Thus, the probability of all confidence intervals being simultaneously correct is not \( 1 - \alpha \), but something less, namely \( 1 - g\alpha \) when the \( t(\frac{\alpha}{2}, N - r) \) distribution is used.

A possible drawback of using the Bonferroni \( t \)-method is that when \( g \) is large, confidence intervals which are too wide for any practical use may be obtained.

## 0.6 Comparison of Bonferroni, Scheffe, and Tukey Methods

The Bonferroni, Scheffe, and Tukey methods all result in the construction of simultaneous confidence intervals which can easily be compared for the set of pairwise comparisons.

If all pairwise comparisons are of interest, the Tukey method gives narrower confidence limits than both the Bonferroni \( t \)-method and the Scheffe method.\(^8\) This implies that more comparisons are significant with the Tukey interval, since if zero is not in the Scheffe interval, or Bonferroni interval, then zero is not in the Tukey intervals. In other words, the Tukey procedure is more powerful than the Scheffe and Bonferroni methods.

If not all pairwise comparisons are of interest, then the Bonferroni method may sometimes lead to narrower confidence intervals than the Tukey method. In addition, the Bonferroni method will produce a smaller interval than the Scheffe method when the number of contrasts to be estimated is equal to or less than the number of factor levels.\(^9\) Of course, the three types of confidence intervals may always be computed separately, in order to determine the smallest.

---

\(^8\text{Neter, Wasserman, and Kutner, Applied Linear Statistical Models, p. 587,589}\)

\(^9\text{Neter, Wasserman, and Kutner, Applied Linear Statistical Models, p. 589}\)
0.7 Dunnett’s Test

In many experiments a control group is utilized. This often leads to the problem of comparing each of the treatments to the control group. Dunnett’s test gives us a method for handling these comparisons. The benefit of using this procedure over the previously mentioned methods, is that increased power is achieved.

Suppose \( Y_{ij} \), \( i = 0, \ldots, r \) and \( j = 1, \ldots, n \) are \( r + 1 \) independent samples of size \( n \) from normal(\( \mu_i, \sigma^2 \)) populations. Let \( \mu_0 \) be the mean for the control group, then Dunnett’s test declares a treatment mean \( \mu_i \) to be larger than the control group if

\[
|\bar{Y}_i - \bar{Y}_0| \geq d(\alpha, r, (r + 1)(n + 1))s(\frac{2}{n})^{\frac{1}{2}}
\]

where \( \bar{Y}_0 \) is the mean for the control group, \( d(\alpha, r, (r + 1)(n + 1)) \) is the \( \alpha \) critical value for the 'many-one t-statistic'\(^{10}\), and

\[
s^2 = \frac{1}{(r + 1)(n - 1)} \sum (Y_{ij} - \bar{Y}_i)^2.
\]

Dunnett’s test guarantees that the experimentwise error rate is not greater than the specified \( \alpha \).\(^{11}\)

\(^{10}\)Miller Simultaneous Statistical Inference, pp. 76-79
\(^{11}\)SAS Stat Users Manual p. 954
0.8 Student-Newman-Keuls and Duncan's Multiple Range Tests

As mentioned in the introduction, the Student-Newman-Keuls (SNK) and Duncan's multiple range tests are more powerful and hence less conservative, than the previously described methods. In addition, the algorithm for constructing the tests is rather simple. Since both tests are constructed using the studentized range distribution, they will be described together as in the book by Arnold.12

Both SNK and Duncan's test are used to make pairwise comparisons of \( r \) means, by proceeding in stages.

1. At the first stage the homogeneity of all means is tested by conducting an \( \alpha_r \)-level studentized range test. That is, we check to see if \( \mu(1) \) and \( \mu(r) \) are significantly different. This is where \( \mu(i) \) is the mean associated with the treatment from which \( \bar{Y}_i \) comes, and \( \bar{Y}_1 \leq \ldots \leq \bar{Y}_r \) are the ordered values of the \( \bar{Y}_i \). If homogeneity is accepted then all means would be declared the same and the procedure stops. If homogeneity if rejected then proceed to step two.

2. At the second stage, the homogeneity of each subset of \( r - 1 \) means is tested by using an \( \alpha_{r-1} \)-level studentized range test. This means we are testing if

(a) \( \mu(r-1) \) and \( \mu(1) \) are significantly different
(b) \( \mu(r) \) and \( \mu(2) \) are significantly different.

If homogeneity is accepted for both (a) and (b), then the only significant different means are declared to be \( \mu(1) \) and \( \mu(r) \), and the procedure stops.

On the other hand, if homogeneity is rejected in at least one of (a) or (b), then we proceed to stage three

3. If we have declared only \( \mu(r-1) \) and \( \mu(1) \) to be different in stage two, then the homogeneity of \( \mu(r-2) \) and \( \mu(1) \) is tested with an \( \alpha_{r-2} \)-level studentized range test.

---

12Arnold, The Theory of Linear Models and Multivariate Analysis, pp. 190-191
BIBLIOGRAPHY


