Multiple comparisons: philosophies and illustrations

DOUGLAS CURRAN-EVERETT
Departments of Preventive Medicine and Biometrics and of Physiology and Biophysics, School of Medicine, University of Colorado Health Sciences Center, Denver, Colorado 80262

Curran-Everett, Douglas. Multiple comparisons: philosophies and illustrations. Am J Physiol Regulatory Integrative Comp Physiol 279: R1–R8, 2000.—Statistical procedures underpin the process of scientific discovery. As researchers, one way we use these procedures is to test the validity of a null hypothesis. Often, we test the validity of more than one null hypothesis. If we fail to use an appropriate procedure to account for this multiplicity, then we are more likely to reach a wrong scientific conclusion—we are more likely to make a mistake. In physiology, experiments that involve multiple comparisons are common: of the original articles published in 1997 by the American Physiological Society, ~40% cite a multiple comparison procedure. In this review, I demonstrate the statistical issue embedded in multiple comparisons, and I summarize the philosophies of handling this issue. I also illustrate the three procedures—Newman-Keuls, Bonferroni, least significant difference—cited most often in my literature review; each of these procedures is of limited practical value. Last, I demonstrate the false discovery rate procedure, a promising development in multiple comparisons. The false discovery rate procedure may be the best practical solution to the problems of multiple comparisons that exist within physiology and other scientific disciplines.

Bonferroni inequality, false discovery rate, least significant difference, Newman-Keuls, statistics

STATISTICAL PROCEDURES are inherent to scientific discovery. As researchers, we use these procedures for two main reasons: to obtain point and interval estimates about the value of a population parameter, and to test the validity of a null hypothesis (5). Point and interval estimates emphasize the magnitude and uncertainty of the experimental results. The test of a null hypothesis helps guard against an unwarranted scientific conclusion, or it helps argue for a real experimental effect (18). When more than one hypothesis is tested—when multiple comparisons are made—the validity of our scientific conclusions may be weakened if we fail to use an appropriate multiple comparison procedure (6, 8, 11, 14, 19, 20).

In studies published recently by the American Physiological Society (APS), the citation of a multiple comparison procedure is common (Table 1). This finding raises an important question: do physiologists understand the philosophies and assumptions behind competing multiple comparison procedures? This question is relevant for three reasons: there are many procedures available, textbooks of statistics (for example, Refs. 1, 13, and 18) provide little more than a cursory description of the procedures themselves, and there can be several solutions to the problem created by multiple comparisons.

In this paper, I summarize the statistical issue embedded in multiple comparisons, and I review the philosophies of handling this issue. Then, I illustrate the three procedures—Newman-Keuls, Bonferroni, least significant difference—cited most often in my literature review. Last, I review the false discovery rate procedure, a promising development in multiple comparisons.

Glossary

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>Error rate for a single comparison</td>
</tr>
<tr>
<td>$\alpha_F$</td>
<td>Error rate for a family of $k$ comparisons</td>
</tr>
<tr>
<td>$H_0$</td>
<td>Null hypothesis</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Population mean</td>
</tr>
<tr>
<td>$P$</td>
<td>Achieved significance level</td>
</tr>
<tr>
<td>$\Pr[A]$</td>
<td>Probability of event $A$</td>
</tr>
</tbody>
</table>
To test a null hypothesis, we must formulate the hypothesis beforehand. Then, using data collected during the experiment, we must compute the observed value $T$ of some test statistic. Last, we must compare the observed value $T$ to a critical value $T^*$, chosen from the distribution of the test statistic that is based on the null hypothesis. If $T$ is more extreme than $T^*$, then that is surprising if the null hypothesis is true, and we are entitled to become skeptical about the scientific validity of the null hypothesis.

Suppose we want to assess renal blood flow in two independent samples. If our objective is to compare the underlying population means, $\mu_1$ and $\mu_2$, then our null and alternative hypotheses, $H_0$ and $H_1$, are

$H_0$: $\mu_1 = \mu_2$

$H_1$: $\mu_1 \neq \mu_2$

The probability that we reject $H_0$ given that $H_0$ is true is the error rate $\alpha$. We can use mathematical notation\(^1\) to write this statement as

$$\Pr(\text{reject } H_0 | H_0 \text{ is true}) = \alpha \quad (1)$$

Note that the critical value $T^*$ is the $100(1 - \alpha/2)$th percentile from the distribution of the test statistic given that the null hypothesis is true. Equation 1 can be rewritten as

$$1 - \Pr(\text{fail to reject } H_0 | H_0 \text{ is true}) = 1 - (1 - \alpha) \Rightarrow \alpha \quad (2)$$

**Multiple comparisons.** Suppose we want to assess renal blood flow in three independent samples.\(^2\) In this setting, there are three alternative hypotheses, $H_1 - H_3$, that correspond to the comparisons among population means:

$H_0$: $\mu_1 = \mu_2 = \mu_3$

$H_1$: $\mu_1 \neq \mu_2$

$H_2$: $\mu_1 \neq \mu_3$

$H_3$: $\mu_2 \neq \mu_3$

Associated with each of these comparisons is an error rate of magnitude $\alpha$. If the three comparisons are considered to be a family, then the family will have an error rate $\alpha_\mathcal{F}$, where $\alpha_\mathcal{F} > \alpha$. As a result, it is more likely that a true null hypothesis will be rejected erroneously. This is the statistical issue that lies at the heart of multiple comparison procedures.

To see why this issue warrants our attention, imagine that each of $k$ independent comparisons is tested at an error rate of $\alpha$. Assume that the underlying populations are identical and that each of the $k$ null hypotheses is true. What is $\alpha_\mathcal{F}$, the probability that at least one of the $k$ comparisons will reject a true null hypothesis? As in Eq. 2, the probability of rejecting at least one $H_0$ given that all $H_0$ are true can be written

$$1 - \Pr(\text{fail to reject all } H_0 | \text{all } k H_0 \text{ are true}) = 1 - (1 - \alpha)^k = \alpha_\mathcal{F}$$

For a single comparison, $\alpha_\mathcal{F} = \alpha$. When the number of comparisons increases, $\alpha$ remains constant, but $\alpha_\mathcal{F}$
increases. For example, if \( \alpha = 0.05 \), then for \( k = 1, 2, 3, 4, 5, \ldots, 10 \),

\[
\begin{array}{cccccccc}
  k & 1 & 2 & 3 & 4 & 5 & \ldots & 10 \\
  \alpha & 0.05 & 0.10 & 0.14 & 0.19 & 0.23 & \ldots & 0.40 \\
\end{array}
\]

For \( k = 10 \) comparisons, there is a 40% chance that we will reject erroneously at least one true null hypothesis.

**Misguided multiple comparisons.** In many of the studies tallied in Table 1, a multiple comparison procedure was used to analyze several groups of observations made on the same subjects. In general, this use of a multiple comparison procedure is misguided: most procedures assume that the groups are independent, but repeated observations on a subject, for example, observations made during baseline and then during several periods after some intervention, create correlation among the groups (9). As a result, the true error variability is underestimated, and the observed values for the standard deviations of the group means underestimate the true variabilities (9). When most multiple comparison procedures are used to analyze groups of repeated observations, the outcome will be an inflated number of statistically significant differences among the group means (see APPENDIX).

**PHILOSOPHIES ABOUT MULTIPLE COMPARISONS**

**Would you tell me, please, which way I ought to go from here?**—Alice

**That depends a good deal on where you want to get to.**—The Cat

L. Carroll in *Alice’s Adventures in Wonderland* (1865)

When we decide the validity of a single comparison, we can make a mistake: we can reject a true null hypothesis, or we can fail to reject a false null hypothesis. When we decide the validity of \( k \) comparisons—this happens in most experiments—we are more likely to reject a true null hypothesis. The challenge for any multiple comparison procedure is to satisfy two conflicting requirements: reduce the risk that we reject a true null hypothesis but maintain the likelihood that we detect an experimental effect if it exists (7, 12, 17). The relative importance assigned to these requirements has produced opposing philosophies about how to handle the issue of multiple comparisons.

**Focus on individual comparisons.** Proponents of this philosophy argue it is sufficient to control the single comparison error rate \( \alpha \), the probability that we reject a true null hypothesis. They base this philosophy on the assumption that most scientific comparisons are preplanned (2, 15, 16). This assumption is naive and unrealistic: many experimental effects are discovered only after an investigator explores—rummages through—the data.

**Control for multiple comparisons.** In general, physiologists examine the impact of an intervention on a set—a family—of related comparisons: for example, the impact of some drug on renal blood flow and urinary excretion of hormones and electrolytes, or a series of paired comparisons among several groups of observations. In these situations, we base our scientific conclusions on a family of comparisons: that is, multiple comparisons considered as a single entity. As a result, it is not the single comparison error rate \( \alpha \) that we must control but the family error rate \( \alpha_{F} \), the probability that we reject at least one true null hypothesis in the family of comparisons (7, 8, 11–13, 17, 19–20). Multiple comparison procedures provide control of the family error rate \( \alpha_{F} \).

**THE GENERAL STRATEGY**

Most multiple comparison procedures use the same basic strategy: to make inferences about the population means for two groups, \( \mu_{1} \) and \( \mu_{2} \), they compare the magnitude of the difference between the sample means \( \bar{y}_{1} \) and \( \bar{y}_{2} \) to a critical difference \( \Delta \bar{y}^{*} \). If

\[
|\bar{y}_{1} - \bar{y}_{2}| > \Delta \bar{y}^{*}
\]

where

\[
\Delta \bar{y}^{*} = c \cdot SE(u)
\]

and where \( SE(u) \) is the standard error of the quantity \( u \), then that is statistical evidence that \( \mu_{1} \neq \mu_{2} \). Procedures differ in the statistics substituted for the coefficient \( c \) and the quantity \( u \). Table 2 lists the statistics for the Newman-Keuls, Bonferroni, and least significant difference tests.

**SIMULATED SAMPLE OBSERVATIONS**

An article published recently in the Journal provides an ideal framework with which to illustrate multiple comparison procedures. In the experiment, Koch et al. (10) explored the heritability of running endurance, measured as distance run, in rats. I used the observed sample statistics from 10 experimental groups (Fig. 1) as the empirical foundation for the simulated sample observations.³

This is how I generated the simulated sample observations—the data. Let the random variable \( Y_{j} \) represent the distance run by a rat in group \( j \), where \( j = 1, 2, \ldots, 10 \). Assume that each \( Y_{j} \) is distributed normally with mean \( \mu_{j} \) and variance \( \sigma_{yj}^{2} \)

\[
Y_{j} \sim N(\mu_{j}, \sigma_{yj}^{2})
\]

---

³ Statistical calculations and exercises were executed using SAS Release 6.12 (SAS Institute, Cary, NC, 1996).

**Table 2. Calculation of the critical difference between sample means, \( \Delta \bar{y}^{*} \)**

<table>
<thead>
<tr>
<th>Procedure</th>
<th>( \alpha_{F} )</th>
<th>( c )</th>
<th>( u )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newman-Keuls</td>
<td>( \alpha )</td>
<td>( \alpha^{n}_{m} )</td>
<td>( \bar{y} )</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>( \alpha )</td>
<td>( \alpha^{n}_{m} )</td>
<td>( \bar{y} )</td>
</tr>
<tr>
<td>Least significant</td>
<td>( \alpha )</td>
<td>( \alpha^{n}_{m} )</td>
<td>( \bar{y} )</td>
</tr>
</tbody>
</table>

For some family error rate \( \alpha_{F} \), the critical difference \( \Delta \bar{y}^{*} \) is \( \Delta \bar{y}^{*} = c \cdot SE(u) \) (Eq 3). The subsequent sections that summarize these multiple comparison procedures detail the quantities for \( \alpha_{F} \) and for the statistics \( c \) and \( u \).
I estimated each $\mu_j$ and $\sigma_j$ using approximate values for the observed group means and standard deviations (see Ref. 10, Tables 1 and 2). For simplicity, I limited each sample to 10 observations. One set of 10 simulated samples is listed in Table 3. For the rest of the review, I use the resulting sample means

$$\bar{y}_1 = 474, \bar{y}_2 = 291, \ldots, \bar{y}_{10} = 612$$

and the resulting sample standard deviations

$$s_1 = 100, s_2 = 102, \ldots, s_{10} = 65$$

as the basis for my illustration of specific multiple comparison procedures.

**NEWMAN-KEULS PROCEDURE**

The Newman-Keuls procedure is a multiple range test that compares the underlying population means of $r$ experimental groups. That is, it evaluates the null hypothesis

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_r \quad (4)$$

The procedure sets the family error rate $\alpha_{gr}$ at $\alpha$, the single comparison error rate, by using studentized range distributions to calculate critical differences (see Eq. 5).

Another multiple range test is the Duncan procedure. It is only the specification of $\alpha_{gr}$ that differentiates the method of Duncan from that of Newman-Keuls. The Duncan family error rate is $\alpha_{gr} = 1 - (1 - \alpha)^{m-1}$, where $m$ is the number of means being compared. The Duncan multiple range test is a noted ancestor of modern multiple comparison procedures, but because $\alpha_{gr}$ grows with $m$, the test violates a basic tenet of multiple comparisons: the control of the single comparison error rate, by using studentized range distributions to calculate critical differences (see Ref. 12, p. 87–89).

**The example.** To make inferences about the equality of two population means, $\mu_r$ and $\mu_j$, the Newman-Keuls procedure uses the critical difference $\Delta \bar{y}_{gr}$, defined as

$$\Delta \bar{y}_{gr} = q_{a,
u} \cdot \text{SE}(\bar{y}) \quad (5)$$

In Eq. 5, the coefficient $q_{a,
u}$ is the $100[1 - \alpha_{gr}]$th percentile of a studentized range distribution with $m$ means and $\nu$ degrees of freedom, and $\text{SE}(\bar{y})$ is the standard error of the sample mean. Using the pooled sample variance $s^2 = 6,883$ (see Table 3), the standard error of the sample mean is estimated as

$$\text{SE}(\bar{y}) = s/\sqrt{n} = 83/\sqrt{10} = 26.2$$

Suppose we define $\alpha_{gr} = 0.05$. In this simulated experiment, there are $r = 90$ degrees of freedom (see Table 3). Because there can be groups of $m = 2, 3, \ldots$, 10 consecutive sample means, there are nine critical differences to be calculated using Eq. 5 (Table 4).

A simple graphical technique can communicate the inferences based on these critical differences. First, we list the sample means in ascending order (see Table 3).

---

**Table 3. Simulated sample observations and derived sample statistics**

<table>
<thead>
<tr>
<th>Group $j$</th>
<th>$\mu_j$</th>
<th>$\sigma_j$</th>
<th>Sample Observations $y_1, y_2, \ldots, y_{10}$</th>
<th>$\bar{y}_j$</th>
<th>$s_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>450</td>
<td>100</td>
<td>501, 619, 382, 502, 480, 396, 269, 543, 547, 501</td>
<td>474</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>325</td>
<td>100</td>
<td>475, 244, 351, 155, 267, 181, 334, 296, 200, 403</td>
<td>291</td>
<td>102</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
<td>100</td>
<td>462, 450, 571, 415, 631, 361, 467, 503, 554, 476</td>
<td>487</td>
<td>75</td>
</tr>
<tr>
<td>4</td>
<td>375</td>
<td>100</td>
<td>487, 356, 498, 336, 489, 411, 248, 369, 423, 423</td>
<td>404</td>
<td>79</td>
</tr>
<tr>
<td>5</td>
<td>650</td>
<td>100</td>
<td>591, 700, 495, 579, 542, 627, 748, 658, 586, 797</td>
<td>632</td>
<td>94</td>
</tr>
<tr>
<td>6</td>
<td>500</td>
<td>100</td>
<td>578, 589, 543, 443, 461, 444, 478, 513, 565, 412</td>
<td>503</td>
<td>64</td>
</tr>
<tr>
<td>7</td>
<td>375</td>
<td>100</td>
<td>331, 440, 406, 339, 389, 372, 286, 341, 498, 349</td>
<td>373</td>
<td>63</td>
</tr>
<tr>
<td>8</td>
<td>400</td>
<td>100</td>
<td>336, 575, 370, 428, 377, 282, 308, 311, 286, 432</td>
<td>370</td>
<td>90</td>
</tr>
<tr>
<td>9</td>
<td>750</td>
<td>100</td>
<td>683, 658, 684, 808, 698, 853, 922, 806, 789, 801</td>
<td>770</td>
<td>86</td>
</tr>
<tr>
<td>10</td>
<td>575</td>
<td>100</td>
<td>564, 616, 632, 700, 674, 663, 561, 544, 505, 661</td>
<td>612</td>
<td>65</td>
</tr>
</tbody>
</table>

---

*Nearly 6% (18/321) of the reviewed manuscripts that report a multiple comparison procedure used the Duncan procedure.*
In fact, for each group of \( m \) consecutive means, progressing from largest to smallest \( m \), we compare the magnitude of the \( m \)-mean range, \( \bar{y}_q - \bar{y}_\ell \), to its corresponding critical difference \( \Delta y^*_m \). If
\[
\bar{y}_q - \bar{y}_\ell \leq \Delta y^*_m
\]
then we underline the group of \( m \) means: we are unable to discriminate among them. If
\[
\bar{y}_q - \bar{y}_\ell > \Delta y^*_m
\]
then we draw no line: we have identified at least one difference. At the end of this process, it is only those means that remain unconnected that we can discriminate statistically.

To illustrate this technique, we begin with \( m = 10 \). The initial step is
\[
770 - 291 = 479 > 120, \text{ draw no line}
\]
In fact, for \( m = 9, 8, \ldots, 4 \), \( \bar{y}_q - \bar{y}_\ell > \Delta y^*_m \), therefore draw no lines.

The next step is to evaluate groups of \( m = 3 \) consecutive means
\[
770 - 612 = 158 > 88, \text{ draw no line;}
\]
\[
632 - 503 = 129 > 88, \text{ draw no line;}
\]
\[
612 - 474 = 125 > 88, \text{ draw no line;}
\]
\[
503 - 474 = 29 > 88, \text{ underline;}
\]
\[
487 - 404 = 83 > 88, \text{ underline;}
\]
\[
474 - 373 = 101 > 88, \text{ draw no line;}
\]
\[
404 - 370 = 34 < 88, \text{ underline;}
\]
\[
373 - 291 = 82 < 88, \text{ underline}
\]
The final step is to evaluate pairs \( (m = 2) \) of adjacent means
\[
770 - 632 = 138 > 74, \text{ draw no line;}
\]
\[
632 - 612 = 20 < 74, \text{ underline;}
\]
\[
612 - 503 = 109 > 74, \text{ draw no line}
\]
At this point, we can stop: all remaining pairs of consecutive means were underlined in the preceding step, when \( m = 3 \).

The Newman-Keuls procedure leads to these conclusions about the 10 sample means
\[
\begin{array}{cccccccccc}
\text{Group } j & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
\bar{y}_j & 291 & 370 & 373 & 404 & 474 & 487 & 503 & 612 & 632 & 770
\end{array}
\]
These are examples of inferences based on this data graphic: \( \mu_2 \) resembles \( \mu_8 \) and \( \mu_7 \) but differs from \( \mu_4 \), \( \mu_1 \), \ldots, \( \mu_9 \); and \( \mu_9 \) differs from all other means. Table 5 lists the inferences for the 16 preplanned group comparisons.

**Practical considerations.** The Newman-Keuls procedure evaluates all \( r(r - 1)/2 \) paired comparisons among \( r \) sample means from a balanced design. The test assumes the \( r \) means are independent and are based on identical numbers of observations (Ref. 12, p. 86). When it compares more than three means, the Newman-Keuls procedure no longer caps the family

<table>
<thead>
<tr>
<th>( m )</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q^*_{m,v} )</td>
<td>2.81</td>
<td>3.37</td>
<td>3.70</td>
<td>3.94</td>
<td>4.12</td>
<td>4.27</td>
<td>4.39</td>
<td>4.50</td>
<td>4.59</td>
</tr>
<tr>
<td>( \Delta y^*_m )</td>
<td>74</td>
<td>88</td>
<td>97</td>
<td>103</td>
<td>108</td>
<td>112</td>
<td>115</td>
<td>118</td>
<td>120</td>
</tr>
</tbody>
</table>

\( q^*_{m,v} \), 100\(1 - \alpha\)th percentile from a studentized range distribution with \( m \) means and \( v \) degrees of freedom; \( \Delta y^*_m \), critical difference for \( m \) consecutive sample means (Eq. 5).

### Table 5. Statistical inferences based on preplanned group comparisons

| Preplanned Comparisons | Statistical Inferences about the Population Means |
|---|---|---|---|---|---|
| \( t \): Null hypothesis \( H^*_0 \) | Newman-Keuls | Bonferroni | LSD | False discovery rate |
| Female rats |
| 1: \( \mu_2 = \mu_3 = \mu_4 \) | 3 | \( \mu_7 \mu_1 \mu_3 \) | \( \mu_7 \mu_4 \mu_8 \) | \( \mu_7 \mu_5 \mu_9 \) | \( \mu_7 \mu_6 \mu_{10} \) | \( \mu_7 \mu_7 \mu_{11} \) | \( \mu_7 \mu_8 \mu_{12} \) |
| 2: \( \mu_2 = \mu_4 = \mu_6 \) | 3 | \( \mu_1 \mu_3 \mu_8 \) | \( \mu_1 \mu_4 \mu_{10} \) | \( \mu_1 \mu_5 \mu_{12} \) | \( \mu_1 \mu_6 \mu_{14} \) | \( \mu_1 \mu_7 \mu_{16} \) | \( \mu_1 \mu_8 \mu_{18} \) |
| Male rats |
| 5: \( \mu_5 = \mu_6 = \mu_7 \) | 3 | \( \mu_2 \mu_3 \mu_4 \) | \( \mu_2 \mu_5 \mu_{10} \) | \( \mu_2 \mu_6 \mu_{14} \) | \( \mu_2 \mu_7 \mu_{18} \) | \( \mu_2 \mu_8 \mu_{22} \) | \( \mu_2 \mu_9 \mu_{26} \) |
| 6: \( \mu_{10} = \mu_6 = \mu_2 \) | 3 | \( \mu_7 \mu_8 \mu_{10} \) | \( \mu_7 \mu_9 \mu_{14} \) | \( \mu_7 \mu_{10} \mu_{18} \) | \( \mu_7 \mu_{12} \mu_{22} \) | \( \mu_7 \mu_{13} \mu_{26} \) | \( \mu_7 \mu_{14} \mu_{30} \) |
| 7: \( \mu_{10} = \mu_8 \) | 1 | \( \mu_8 \mu_{10} \mu_{12} \) | \( \mu_8 \mu_{10} \mu_{10} \) | \( \mu_8 \mu_{10} \mu_{10} \) | \( \mu_8 \mu_{10} \mu_{10} \) | \( \mu_8 \mu_{10} \mu_{10} \) | \( \mu_8 \mu_{10} \mu_{10} \) |
| 8: \( \mu_6 = \mu_4 \) | 1 | \( \mu_5 \mu_6 \mu_8 \) | \( \mu_5 \mu_6 \mu_8 \) | \( \mu_5 \mu_6 \mu_8 \) | \( \mu_5 \mu_6 \mu_8 \) | \( \mu_5 \mu_6 \mu_8 \) | \( \mu_5 \mu_6 \mu_8 \) |

\( k_o \), number of comparisons associated with the null hypothesis; LSD, least significant difference. For each multiple comparison procedure, the relative ordering of the population means matches that of the sample means because the sample mean \( \bar{y}_j \) estimates the population mean \( \mu_i \); that is, because \( \bar{y}_j = \bar{\mu}_j \). Underlined population means cannot be discriminated statistically. Note that the Bonferroni inequality fails to detect several differences between means that the other procedures identify.
error rate \( \alpha_f \) at \( \alpha \); instead, \( \alpha_f > \alpha \) (Ref. 8, p. 127). For this reason, the Newman-Keuls procedure is of limited value for multiple comparisons.

**BONFERRONI PROCEDURE**

The Bonferroni inequality is a probability inequality that does control the family error rate \( \alpha_f \). For a family of \( k \) comparisons, the Bonferroni inequality defines the upper bound of the family error rate to be

\[
\alpha_f = 1 - (1 - \alpha)^k = k \cdot \alpha
\]

where \( \alpha \) is the error rate for each comparison. In other words, the inequality assigns an error rate of \( \alpha_f / k \) to each comparison within the family. Because \( \alpha \) can vary among comparisons, the general expression for the family error rate is

\[
\alpha_f = \alpha_1 + \alpha_2 + \cdots + \alpha_k
\]

The example. To make inferences about the equality of two population means, \( \mu_k \) and \( \mu_{k'} \), the Bonferroni procedure relies on the critical difference \( \Delta y^* \), defined as

\[
\Delta y^* = t_{a_1/2, \nu} \cdot SE(\bar{y}_k - \bar{y}_{k'})
\]

In Eq. 6, the coefficient \( t_{a_1/2, \nu} \) is the 100\([1 - (\alpha_f/2)]\)th percentile from a \( t \) distribution with \( \nu \) degrees of freedom, and \( SE(\bar{y}_k - \bar{y}_{k'}) \) is the standard error of the difference between the sample means.

If we define \( \alpha_f = 0.05 \), then for each of the 16 preplanned comparisons listed in Table 5

\[
\alpha = \alpha_f / k = 0.05 / 16 = 0.0031
\]

Therefore, because there are \( \nu = 90 \) degrees of freedom (see Table 3), \( t_{a_1/2, \nu} = 3.04 \). Using the pooled sample variance \( s^2 = 6,883 \), the standard error of the difference between sample means is estimated as

\[
SE(\bar{y}_k - \bar{y}_{k'}) = \sqrt{(s^2 + s^2)/n} = 37.1
\]

By virtue of Eq. 6, the resulting critical difference for the Bonferroni procedure is

\[
\Delta y^* = 3.04 \times 37.1 = 113
\]

Therefore, the Bonferroni procedure leads to these conclusions about the 10 sample means

<table>
<thead>
<tr>
<th>Group</th>
<th>2</th>
<th>8</th>
<th>7</th>
<th>4</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>10</th>
<th>5</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{y} )</td>
<td>291</td>
<td>370</td>
<td>373</td>
<td>404</td>
<td>474</td>
<td>487</td>
<td>503</td>
<td>612</td>
<td>632</td>
<td>770</td>
</tr>
</tbody>
</table>

Table 5 lists the resulting inferences for the 16 preplanned group comparisons.

Practical considerations. Although it is not a multiple comparison procedure per se, the Bonferroni inequality can be used for multiple comparison problems. The technique is valid regardless of whether the \( r \) sample means are independent or correlated (Ref. 12, p. 67). The Bonferroni inequality is appealing because it is versatile and simple. Unfortunately, its appeal is diminished by the strict protection of the single comparison error rate \( \alpha \). As a consequence, the Bonferroni inequality is conservative: it will be unable to detect some of the actual differences among a family of \( k \) comparisons (see Table 5).

**LEAST SIGNIFICANT DIFFERENCE PROCEDURE**

The least significant difference (LSD) procedure, developed by Sir R. A. Fisher, preceded the Newman-Keuls multiple range test. Like the Newman-Keuls test, the LSD procedure compares the underlying population means of \( r \) experimental groups (see Eq. 4), and it sets the family error rate \( \alpha_f \) at the single comparison error rate \( \alpha \).

The example. To make inferences about the equality of two population means, \( \mu_k \) and \( \mu_{k'} \), the LSD procedure uses the critical difference \( \Delta y^* \), defined as

\[
\Delta y^* = t_{a_1/2, \nu} \cdot SE(\bar{y}_k - \bar{y}_{k'})
\]

In Eq. 8, the coefficient \( t_{a_1/2, \nu} \) is the 100\([1 - (\alpha_f/2)]\)th percentile from a \( t \) distribution with \( \nu \) degrees of freedom, and \( SE(\bar{y}_k - \bar{y}_{k'}) \) is the standard error of the difference between the sample means.

If we define \( \alpha_f = 0.05 \), then because there are \( \nu = 90 \) degrees of freedom (see Table 3), \( t_{a_1/2, \nu} = 1.99 \). As shown in Eq. 7, \( SE(\bar{y}_k - \bar{y}_{k'}) = 37.1 \). Therefore, by virtue of Eq. 8, the resulting critical difference for the LSD procedure is

\[
\Delta y^* = 1.99 \times 37.1 = 74
\]

The LSD procedure leads to these conclusions about the 10 sample means

<table>
<thead>
<tr>
<th>Group</th>
<th>2</th>
<th>8</th>
<th>7</th>
<th>4</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>10</th>
<th>5</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{y} )</td>
<td>291</td>
<td>370</td>
<td>373</td>
<td>404</td>
<td>474</td>
<td>487</td>
<td>503</td>
<td>612</td>
<td>632</td>
<td>770</td>
</tr>
</tbody>
</table>

Table 5 lists the resulting inferences for the 16 preplanned group comparisons.

Practical considerations. The LSD procedure evaluates all \( r(r - 1)/2 \) paired comparisons among \( r \) sample means. In its protected form, the procedure is done only if a preliminary analysis of variance is statistically significant (18). When it compares more than three means, the LSD procedure fails to maintain the family error rate \( \alpha_f \) at \( \alpha \) (Ref. 8, p. 139). The solution to this problem is to replace \( t_{a_1/2, \nu} \) in Eq. 8 with a percentile from a studentized range distribution: \( q_{r-1, \nu} \) (Ref. 8, p. 139) or \( a_{r, \nu} \) (Ref. 12, p. 92).

**FALSE DISCOVERY RATE PROCEDURE:**

A RECENT DEVELOPMENT

In most experiments, scientists strive to make a discovery: to reject a null hypothesis. When an experiment involves a family of \( k \) comparisons, a scientist is more likely to make a mistaken discovery. The false discovery rate procedure is a promising solution to the...
The problem of multiple comparisons. This procedure controls not the family error rate $\alpha_F$ but the false discovery rate $f_D$, the expected fraction of null hypotheses rejected mistakenly

$$f_D = \frac{\text{number of mistaken } H_0 \text{ rejections}}{\text{total number of } H_0 \text{ rejections}}$$

If all $k$ null hypotheses are true, then $f_D = \alpha_F$; if at least one null hypothesis is not true, then $f_D \leq \alpha_F$ (3). When we define the family error rate $\alpha_F$, we also set an upper bound on the false discovery rate $f_D$. But if we control $f_D$ rather than $\alpha_F$, we gain statistical power, the ability to detect an experimental effect if it exists (3, 4, 22).

The example. Unlike the preceding methods, the false discovery rate procedure operates on achieved significance levels ($P$ values) to make inferences about a family of $k$ comparisons. Let $P_i$ represent the significance level associated with comparison $i$. To execute this procedure, we must complete three steps:

**Step 1.** Order the $k$ comparisons by decreasing magnitude of $P_i$.

**Step 2.** For $i = k, k - 1, \ldots, 1$, calculate the critical significance level $d_i$ as

$$d_i = (i/k) \cdot f_D$$

**Step 3.** If $P_i \leq d_i$, then reject the null hypotheses associated with the remaining $i$ comparisons.

In the simulation, we selected $k = 16$ comparisons of interest. For each comparison, we evaluate the null hypothesis $H_0$: $\mu_i = \mu_q$ by doing a $t$ test. The $P$ values associated with the resulting $t$ statistics vary from 0.723 to 0.001 (Table 6). If we define the false discovery rate $f_D = 0.05$, the magnitude of the family error rate $\alpha_F$ we have been using, then the critical significance level $d_i$ varies from 0.050 to 0.003. In step 3, we declare comparisons 1–14 to be statistically significant (see Table 5).

Table 6. Calculations for the false discovery rate procedure

<table>
<thead>
<tr>
<th>Comparison $i$</th>
<th>$H_i$</th>
<th>$P_i$</th>
<th>$d_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>16: $\mu_3 = \mu_4$</td>
<td>0.723</td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td>15: $\mu_3 = \mu_4$</td>
<td>0.389</td>
<td>0.047</td>
<td></td>
</tr>
<tr>
<td>14: $\mu_3 = \mu_4$</td>
<td>0.034</td>
<td>0.044</td>
<td></td>
</tr>
<tr>
<td>13: $\mu_3 = \mu_4$</td>
<td>0.009</td>
<td>0.041</td>
<td></td>
</tr>
<tr>
<td>12: $\mu_3 = \mu_4$</td>
<td>0.008</td>
<td>0.038</td>
<td></td>
</tr>
<tr>
<td>11: $\mu_3 = \mu_4$</td>
<td>0.004</td>
<td>0.034</td>
<td></td>
</tr>
<tr>
<td>10: $\mu_3 = \mu_4$</td>
<td>0.003</td>
<td>0.031</td>
<td></td>
</tr>
<tr>
<td>9: $\mu_3 = \mu_4$</td>
<td>0.003</td>
<td>0.028</td>
<td></td>
</tr>
<tr>
<td>8: $\mu_3 = \mu_4$</td>
<td>0.001</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td>7: $\mu_3 = \mu_4$</td>
<td>0.001</td>
<td>0.022</td>
<td></td>
</tr>
<tr>
<td>6: $\mu_3 = \mu_4$</td>
<td>0.001</td>
<td>0.019</td>
<td></td>
</tr>
<tr>
<td>5: $\mu_3 = \mu_4$</td>
<td>0.001</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>4: $\mu_3 = \mu_4$</td>
<td>0.001</td>
<td>0.012</td>
<td></td>
</tr>
<tr>
<td>3: $\mu_3 = \mu_4$</td>
<td>0.001</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td>2: $\mu_3 = \mu_4$</td>
<td>0.001</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>1: $\mu_3 = \mu_4$</td>
<td>0.001</td>
<td>0.003</td>
<td></td>
</tr>
</tbody>
</table>

$P_i$ achieved significance level; $d_i$ critical significance level (Eq. 9). For $P_i$ a value of 0.001 denotes $P_i < 0.01$. If $P_i < d_i$, then the remaining $i$ null hypotheses are rejected. Because $P_{14} = 0.034 < d_i$, null hypotheses 14 → 1 are rejected. See Table 5 for a graphical depiction of these numerical results.

**SUMMARY**

We dare not seek a single multiple comparison procedure for all experiments.

Adapted from John W. Tukey (1994)

This remark, written by a pioneer in the area of multiple comparisons, reflects the range of multiple comparison problems that manifest themselves in scientific research. Over the last 50–60 years, statisticians have explored numerous approaches in an effort to address these problems (8, 12). In physiology, as in other disciplines, experiments that involve problems of multiple comparisons are common.

In this review, I have shown that, as researchers, we are more likely to reject a true null hypothesis if we fail to use a multiple comparison procedure when we analyze a family of comparisons. I have also illustrated the three procedures cited most often in APS journals: Newman-Keuls, Bonferroni, and LSD. Unfortunately, each of these is of limited value. In many experimental situations, the Newman-Keuls and LSD procedures fail to control the family error rate, the probability that we reject at least one true null hypothesis. In contrast, the Bonferroni inequality is overly conservative: it fails to detect some of the actual differences that exist within the family.

Finally, I have reviewed the false discovery rate: a versatile, simple, and powerful approach to multiple comparisons. As Tukey suggests, it is perhaps unrealistic to expect that a single multiple comparison procedure will suffice for all situations: a statistical procedure designed specifically for a particular experimental situation will
perform better than a general procedure. Nevertheless, there is growing evidence (4, 22) that the false discovery rate procedure may be the best practical solution to the problems of multiple comparisons that exist within science.

APPENDIX

For all but one of the multiple comparison procedures listed in Table 1, an important assumption is that the $r$ experimental groups are independent (12). In many studies that use these multiple comparison procedures, however, the $r$ groups are not independent. This happens because investigators make repeated observations on each subject: these observations are correlated by virtue of individual biological variability. Therefore, the true error variability is underestimated, and the observed values for the standard deviations are distributed normally.

Assume that $X_1$ and $X_2$ are considered jointly, then the distribution of the variable pair $(X_1, X_2)$ can be envisioned as a bivariate normal distribution. For this distribution, $\sigma_{X_1}$, the standard deviation of the conditional distribution of $X_2$ given that $X_1$ equals a specific value, depends on the correlation $\rho$ between $X_1$ and $X_2$:

$$\sigma_{X_1} = \sigma_2 \sqrt{1 - \rho^2}, \quad \text{where} \quad -1 \leq \rho \leq 1$$

Because repeated observations on a subject are correlated, that is, because $\rho \neq 0$, the standard deviation of the variable measured during a second condition, given the value of the first measurement, is reduced by a factor of $\sqrt{1 - \rho^2}$.

I thank Dr. Steven L. Britton (Department of Physiology and Molecular Medicine, Medical College of Ohio) and colleagues for permission to cite their study.

REFERENCES


---

11The lone exception is the Bonferroni inequality, which allows the $r$ experimental groups to be correlated.