Am J Physiol Regulatory Integrative Comp Physiol
279: R1–R8, 2000.

invited review

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, a promising development in multiple comparisons.

Glossary

α	Error rate for a single comparison
$\alpha_{\mathcal{F}}$	Error rate for a family of k comparisons
H_0	Null hypothesis
μ	Population mean
P	Achieved significance level
$\Pr\{A\}$	Probability of event A

Address for reprint requests and other correspondence: D. Curran-Everett, Department of Preventive Medicine and Biometrics, B-195, University of Colorado Health Sciences Center, 4200 East 9th Ave., Denver, CO 80262 (E-mail: dcurran-@carbon.cudenver.edu).

Table 1. *Manuscripts of APS journals in 1997: use of multiple comparison procedures*

	Manuscripts		Procedures Used, %†			
	n	% Multiple comparisons*	Newman-Keuls	Bonferroni	LSD	Other‡
Am J Physiol						
Cell Physiol	90	20	17	33	11	39
Endocrinol Metab	61	49	30	7	12	52
Gastrointest Liver Physiol	68	37	16	24	12	48
Heart Circ Physiol	136	60	23	21	9	47
Lung Cell Mol Physiol	62	52	17	25	20	38
Regulatory Integrative Comp Physiol	106	56	23	12	16	48
Renal Physiol	44	20	6	39	0	56
J Appl Physiol	106	53	25	20	11	45
J Neurophysiol	125	8	25	40	20	15
All journals taken together	798	40	22	20	12	45

n, number of research manuscripts reviewed; LSD, least significant difference. In 1997, these journals published a total of ~4,000 original articles. Number of articles reviewed represents a 20% sample (selected by systematic random sampling, fixed start) of the articles published by each journal. *Percentage of research manuscripts that report a multiple comparison procedure. †Values represent the % use in those manuscripts that report a multiple comparison procedure. ‡Includes Duncan, Dunnett, Scheffé, Tukey, and unnamed procedures. Roughly 8% (27/321) of manuscripts that report use of a multiple comparison procedure fail to identify the procedure.

$\bar{y}$ Sample mean
$\Delta\bar{y}^*$ Critical difference between two sample means

THE ISSUE EMBEDDED IN MULTIPLE COMPARISONS

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, μ_1 and μ_2, then one pair of null and alternative hypotheses, H_0 and H_1, is

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

The probability that we reject H_0 given that H_0 is true is the error rate α. 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

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

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 α. 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 α. 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, α remains constant, but $\alpha_{\mathcal{F}}$

[1] In comments about my review of statistical concepts (Ref. 5), one referee wrote that my exposition was mathematical and therefore unfriendly. I use mathematics for two reasons: mathematics is one dialect of the language of science, and the precision of mathematical notation simplifies communication and clarifies reasoning. Nevertheless, because I appreciate that readers will have different levels of comfort with mathematics, I integrate the mathematics with text summaries.

[2] For r experimental groups, there are $r(r - 1)/2$ paired comparisons possible.

increases. For example, if $\alpha = 0.05$, then for $k = 1, 2, 3, 4, 5, \ldots, 10$,

k	1	2	3	4	5	. . .	10
$\alpha_{\mathcal{F}}$	0.05	0.10	0.14	0.19	0.23	. . .	0.40

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 α, 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 α that we must control but the family error rate $\alpha_{\mathcal{F}}$, 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 $\alpha_{\mathcal{F}}$.

THE GENERAL STRATEGY

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

$$|\bar{y}_\varphi - \bar{y}_\ell| > \Delta\bar{y}^*$$

where

$$\Delta\bar{y}^* = c \cdot \mathrm{SE}\{u\} \qquad (3)$$

and where SE{u} is the standard error of the quantity u, then that is statistical evidence that $\mu_\ell \neq \mu_\varphi$. 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 μ_j and variance σ_j^2

$$Y_j \sim N(\mu_j, \sigma_j^2)$$

[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,* $\Delta\bar{y}^*$

	$\alpha_{\mathcal{F}}$	c	u
Newman-Keuls	α	$q_{m,\nu}^{\alpha_{\mathcal{F}}}$	$\bar{y}$
Bonferroni	$k\alpha$	$t_{\alpha/2,\nu}$	$\bar{y}_\varphi - \bar{y}_\ell$
Least significant difference	α	$t_{\alpha_{\mathcal{F}}/2,\nu}$	$\bar{y}_\varphi - \bar{y}_\ell$

For some family error rate $\alpha_{\mathcal{F}}$, the critical difference $\Delta\bar{y}^*$ is $\Delta\bar{y}^* = c \cdot \mathrm{SE}\{u\}$ (*Eq 3*). The subsequent sections that summarize these multiple comparison procedures detail the quantities for $\alpha_{\mathcal{F}}$ and for the statistics c and u.

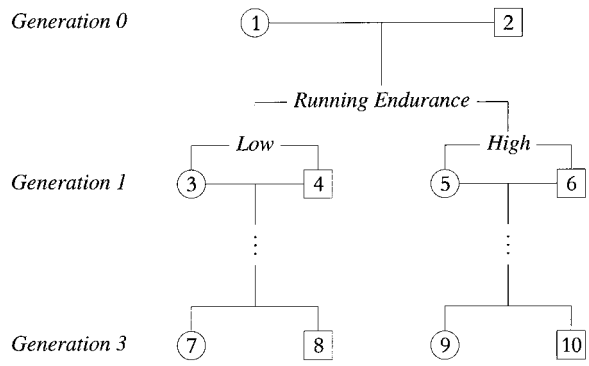

Fig. 1. Experimental *groups 1–10* associated with the simulated sample observations and derived sample statistics listed in Table 3. This diagram is based on the selective breeding procedure described in Ref. 10. The initial generation is *generation 0.* In each generation, the 2 female (○) and 2 male (□) rats at the extremes of observed running endurance were paired and bred to produce the subsequent generation.

I estimated each μ_j and σ_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

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

and the resulting sample standard deviations

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

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

NEWMAN-KEULS PROCEDURE

The Newman-Keuls procedure[4] 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 $\alpha_{\mathcal{F}}$ at α, 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.[5] It is only the specification of $\alpha_{\mathcal{F}}$ that differentiates the method of Duncan from that of Newman-Keuls. The Duncan family error rate is $\alpha_{\mathcal{F}} = 1 - (1 - \alpha)^{m-1}$, 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 $\alpha_{\mathcal{F}}$ grows with m, the test violates a basic tenet of multiple comparisons: the control of $\alpha_{\mathcal{F}}$ despite a large number of comparisons (see Ref. 12, p. 87–89).

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

$$\Delta\bar{y}^*_m = q^{\alpha_{\mathcal{F}}}_{m,\nu} \cdot \mathrm{SE}\{\bar{y}\} \quad (5)$$

In *Eq. 5,* the coefficient $q^{\alpha_{\mathcal{F}}}_{m,\nu}$ is the $100[1 - \alpha_{\mathcal{F}}]$th percentile from a studentized range distribution with m means and ν degrees of freedom, and $\mathrm{SE}\{\bar{y}\}$ 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

$$\mathrm{SE}\{\bar{y}\} = s/\sqrt{n} = 83/\sqrt{10} = 26.2$$

Suppose we define $\alpha_{\mathcal{F}} = 0.05$. In this simulated experiment, there are $\nu = 90$ 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)

[4]This procedure is known also by the name Student-Newman-Keuls.

[5]Nearly 6% (18/321) of the reviewed manuscripts that report a multiple comparison procedure used the Duncan procedure.

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

Group j	μ_j	σ_j	Sample Observations $y_1, y_2, \ldots, y_{10}$	$\bar{y}_j$	s_j
1	450	100	501, 619, 382, 502, 480, 396, 269, 543, 547, 501	474	100
2	325	100	475, 244, 351, 155, 267, 181, 334, 296, 200, 403	291	102
3	500	100	462, 450, 571, 415, 613, 361, 467, 503, 554, 476	487	75
4	375	100	487, 356, 498, 336, 489, 411, 248, 369, 423, 423	404	79
5	650	100	591, 700, 495, 579, 542, 627, 748, 658, 586, 797	632	94
6	500	100	578, 589, 543, 443, 461, 444, 478, 513, 565, 412	503	64
7	375	100	313, 440, 406, 339, 389, 372, 286, 341, 498, 349	373	63
8	400	100	336, 575, 370, 428, 377, 282, 308, 311, 286, 432	370	90
9	750	100	683, 658, 684, 808, 698, 853, 922, 806, 789, 801	770	86
10	575	100	564, 616, 632, 700, 674, 663, 561, 544, 505, 661	612	65

Experimental *groups 1–10* correspond to those depicted in Fig. 1. From each of the 10 populations defined by the mean μ_j and standard deviation σ_j, we draw 10 independent sample observations that represent distance run (in m). Each sample mean $\bar{y}_j$ and sample standard deviation s_j estimate the corresponding population mean μ_j and population standard deviation σ_j. Because there are $r = 10$ groups, each with $n = 10$ observations in each group, there are $\nu = r \cdot (n - 1)$ degrees of freedom. In this simulation, the pooled sample variance $s^2 = 6{,}883$. These values are used to calculate the critical difference $\Delta\bar{y}^*$ for the Newman-Keuls, Bonferroni, and least significant difference procedures.

83^2 = 6883

Table 4. *Critical differences for the Newman-Keuls procedure*

	m, Number of Means Being Compared								
	2	3	4	5	6	7	8	9	10
$q^{\alpha_{\mathcal{F}}}_{m,\nu}$	2.81	3.37	3.70	3.94	4.12	4.27	4.39	4.50	4.59
$\Delta\bar{y}^*_m$	74	88	97	103	108	112	115	118	120

$q^{\alpha_{\mathcal{F}}}_{m,\nu}$, 100[1 − $\alpha_{\mathcal{F}}$]th percentile from a studentized range distribution with m means and ν degrees of freedom; $\Delta\bar{y}^*_m$, critical difference for m consecutive sample means (*Eq. 5*).

Group j	2	8	7	4	1	3	6	10	5	9
$\bar{y}_j$	291	370	373	404	474	487	503	612	632	770

Then, for each group of m consecutive means, progressing from largest to smallest m, we compare the magnitude of the m-mean range, $\bar{y}_\varphi - \bar{y}_\ell$, to its corresponding critical difference $\Delta\bar{y}^*_m$. If

$$\bar{y}_\varphi - \bar{y}_\ell \leq \Delta\bar{y}^*_m$$

then we underline the group of m means: we are unable to discriminate among them. If

$$\bar{y}_\varphi - \bar{y}_\ell > \Delta\bar{y}^*_m$$

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 $m = 9, 8, \ldots, 4, \bar{y}_\varphi - \bar{y}_\ell > \Delta\bar{y}^*_m$, therefore draw no lines.

The next step is to evaluate groups of $m = 3$ consecutive means

$$770 - 612 = 158 > 88, \text{ draw no line;}$$
$$632 - 503 = 129 > 88, \text{ draw no line;}$$
$$612 - 487 = 125 > 88, \text{ draw no line;}$$
$$503 - 474 = 29 < 88, \text{ underline;}$$
$$487 - 404 = 83 < 88, \text{ underline;}$$
$$474 - 373 = 101 > 88, \text{ draw no line;}$$
$$404 - 370 = 34 < 88, \text{ underline;}$$
$$373 - 291 = 82 < 88, \text{ underline}$$

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

Group j	2	8	7	4	1	3	6	10	5	9
$\bar{y}_j$	291	370	373	404	474	487	503	612	632	770

These are examples of inferences based on this data graphic: μ_2 resembles μ_8 and μ_7 but differs from μ_4, $\mu_1, \ldots, \mu_9$; and μ_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 5. *Statistical inferences based on preplanned group comparisons*

Preplanned Comparisons		Statistical Inferences about the Population Means			
i: Null hypothesis H^i_0	k_0	Newman-Keuls	Bonferroni	LSD	False discovery rate
Female rats					
1: $\mu_7 = \mu_3 = \mu_1$	3	μ_7 $\underline{\mu_1}$ $\underline{\mu_3}$	$\underline{\mu_7}$ $\underline{\mu_1}$ $\underline{\mu_3}$	μ_7 $\underline{\mu_1}$ $\underline{\mu_3}$	μ_7 $\underline{\mu_1}$ $\underline{\mu_3}$
2: $\mu_9 = \mu_5 = \mu_1$	3	μ_1 $\underline{\mu_5}$ $\underline{\mu_9}$	$\underline{\mu_1}$ $\underline{\mu_5}$ $\underline{\mu_9}$	μ_1 $\underline{\mu_5}$ $\underline{\mu_9}$	μ_1 $\underline{\mu_5}$ $\underline{\mu_9}$
3: $\mu_9 = \mu_7$	1	μ_7 μ_9	μ_7 μ_9	μ_7 μ_9	μ_7 μ_9
4: $\mu_5 = \mu_3$	1	μ_3 μ_5	μ_3 μ_5	μ_3 μ_5	μ_3 μ_5
Male rats					
5: $\mu_8 = \mu_4 = \mu_2$	3	$\underline{\mu_2}$ $\underline{\mu_8}$ $\underline{\mu_4}$	$\underline{\mu_2}$ $\underline{\mu_8}$ $\underline{\mu_4}$	$\underline{\mu_2}$ $\underline{\mu_8}$ $\underline{\mu_4}$	$\underline{\mu_2}$ $\underline{\mu_8}$ $\underline{\mu_4}$
6: $\mu_{10} = \mu_6 = \mu_2$	3	$\underline{\mu_2}$ $\underline{\mu_6}$ $\underline{\mu_{10}}$	$\underline{\mu_2}$ $\underline{\mu_6}$ $\underline{\mu_{10}}$	$\underline{\mu_2}$ $\underline{\mu_6}$ $\underline{\mu_{10}}$	$\underline{\mu_2}$ $\underline{\mu_6}$ $\underline{\mu_{10}}$
7: $\mu_{10} = \mu_8$	1	μ_8 μ_{10}	μ_8 μ_{10}	μ_8 μ_{10}	μ_8 μ_{10}
8: $\mu_6 = \mu_4$	1	μ_4 μ_6	$\underline{\mu_4}$ $\underline{\mu_6}$	μ_4 μ_6	μ_4 μ_6

k_0, 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 $\bar{y}_i$ estimates the population mean μ_i: that is, because $\bar{y}_i = \hat{\mu}_i$. 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 $\alpha_{\mathcal{F}}$ at α; instead, $\alpha_{\mathcal{F}} > \alpha$ (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 $\alpha_{\mathcal{F}}$. For a family of k comparisons, the Bonferroni inequality defines the upper bound of the family error rate to be

$$\alpha_{\mathcal{F}} = 1 - (1 - \alpha)^k = k \cdot \alpha$$

where α is the error rate for each comparison. In other words, the inequality assigns an error rate of $\alpha_{\mathcal{F}}/k$ to each comparison within the family. Because α can vary among comparisons, the general expression for the family error rate is

$$\alpha_{\mathcal{F}} = \alpha_1 + \alpha_2 + \cdots + \alpha_k$$

The example. To make inferences about the equality of two population means, μ_ℓ and μ_φ, the Bonferroni procedure relies on the critical difference $\Delta\bar{y}^*$, defined as

$$\Delta\bar{y}^* = t_{\alpha/2,\nu} \cdot \mathrm{SE}\{\bar{y}_\varphi - \bar{y}_\ell\} \qquad (6)$$

In *Eq. 6,* the coefficient $t_{\alpha/2,\nu}$ is the $100[1 - (\alpha/2)]$th percentile from a t distribution with ν degrees of freedom, and $\mathrm{SE}\{\bar{y}_\varphi - \bar{y}_\ell\}$ is the standard error of the difference between the sample means.

If we define $\alpha_{\mathcal{F}} = 0.05$, then for each of the 16 preplanned comparisons listed in Table 5

$$\alpha = \alpha_{\mathcal{F}}/k = 0.05/16 = 0.0031$$

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

$$\mathrm{SE}\{\bar{y}_\varphi - \bar{y}_\ell\} = \sqrt{(s^2 + s^2)/n} = 37.1 \qquad (7)$$

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

$$\Delta\bar{y}^* = 3.04 \times 37.1 = 113$$

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

Group j	2	8	7	4	1	3	6	10	5	9
$\bar{y}_j$	291	370	373	404	474	487	503	612	632	770

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 α. 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 $\alpha_{\mathcal{F}}$ at the single comparison error rate α.

The example. To make inferences about the equality of two population means, μ_ℓ and μ_φ, the LSD procedure uses the critical difference $\Delta\bar{y}^*$, defined as

$$\Delta\bar{y}^* = t_{\alpha_{\mathcal{F}}/2,\nu} \cdot \mathrm{SE}\{\bar{y}_\varphi - {}_\ell\} \qquad (8)$$

In *Eq. 8,* the coefficient $t_{\alpha_{\mathcal{F}}/2,\nu}$ is the $100[1 - (\alpha_{\mathcal{F}}/2)]$th percentile from a t distribution with ν degrees of freedom, and $\mathrm{SE}\{\bar{y}_\varphi - \bar{y}_\ell\}$ is the standard error of the difference between the sample means.[6]

If we define $\alpha_{\mathcal{F}} = 0.05$, then because there are $\nu = 90$ degrees of freedom (see Table 3), $t_{\alpha_{\mathcal{F}}/2,\nu} = 1.99$. As shown in *Eq. 7,* $\mathrm{SE}\{\bar{y}_\varphi - \bar{y}_\ell\} = 37.1$. Therefore, by virtue of *Eq. 8,* the resulting critical difference for the LSD procedure is

$$\Delta\bar{y}^* = 1.99 \times 37.1 = 74$$

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

Group j	2	8	7	4	1	3	6	10	5	9
$\bar{y}_j$	291	370	373	404	474	487	503	612	632	770

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 $\alpha_{\mathcal{F}}$ at α (Ref. 8, p. 139). The solution to this problem is to replace $t_{\alpha_{\mathcal{F}}/2,\nu}$ in *Eq. 8* with a percentile from a studentized range distribution: $q^{\alpha_{\mathcal{F}}}_{r-1,\nu}$ (Ref. 8, p. 139) or $q^{\alpha_{\mathcal{F}}}_{r,\nu}$ (Ref. 12, p. 92).[7]

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[8] is a promising solution to the

[6] Because $\alpha_{\mathcal{F}} = \alpha$, this critical difference is simply the allowance used to obtain a $100(1 - \alpha)\%$ confidence interval for the difference $\bar{y}_\varphi - \bar{y}_\ell$ (see Ref. 5, *Eq. A2*).

[7] When the latter coefficient is used in *Eq. 8,* the method is called the wholly (or honestly) significant difference procedure.

[8] This procedure is available within SAS Release 6.12 by using the `fdr` option in `Proc MultTest`.

problem of multiple comparisons. This procedure controls not the family error rate $\alpha_{\mathcal{F}}$ but the false discovery rate $f_{\mathcal{F}}$, the expected fraction of null hypotheses rejected mistakenly

$$f_{\mathcal{F}} = \frac{\text{number of mistaken } H_0 \text{ rejections}}{\text{total number of } H_0 \text{ rejections}}$$

If all k null hypotheses are true,[9] then $f_{\mathcal{F}} = \alpha_{\mathcal{F}}$; if at least one null hypothesis is not true, then $f_{\mathcal{F}} \leq \alpha_{\mathcal{F}}$ (3). When we define the family error rate $\alpha_{\mathcal{F}}$, we also set an upper bound on the false discovery rate $f_{\mathcal{F}}$. But if we control $f_{\mathcal{F}}$ rather than $\alpha_{\mathcal{F}}$, 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

$$d_i^* = (i/k) \cdot f_{\mathcal{F}} \qquad (9)$$

Step 3. If $P_i \leq d_i^*$, then reject the null hypotheses associated with the remaining i comparisons.[10]

In the simulation, we selected $k = 16$ comparisons of interest. For each comparison, we evaluate the null hypothesis H_0: $\mu_\ell = \mu_\varphi$ by doing a t test. The P values associated with the resulting t statistics vary from 0.723 $\rightarrow 0.001^-$ (Table 6). If we define the false discovery rate $f_{\mathcal{F}} = 0.05$, the magnitude of the family error rate $\alpha_{\mathcal{F}}$ we have been using, then the critical significance level d_i^* varies from 0.050 $\rightarrow$ 0.003. In *step 3,* we declare *comparisons 1–14* to be statistically significant (see Table 6). Table 5 lists the inferences for all 16 comparisons.

Practical considerations. Because the false discovery rate procedure operates on actual P values, it is quite versatile. For example, the procedure can be employed when a family of k comparisons involves different test statistics such as Student t and Wilcoxon signed rank statistics (3, 4). The false discovery rate procedure is valid when the k comparisons are independent (a sample mean is part of only one comparison) or correlated (a sample mean is part of more than one comparison, as in the example) (3, 4, 22).

The false discovery rate procedure has two important benefits. First, it allows us to make an inference, with $100[1 - (f_{\mathcal{F}}/2)]\%$ confidence, about the direction of a statistical difference (4, 22). For example, because $f_{\mathcal{F}} = 0.05$, we can declare, with 97.5% confidence, that $\mu_2 < \mu_8$ (see Table 6). This is a stronger inference than the simple declaration $\mu_2 \neq \mu_8$ (Ref. 8, p. 27–39). Second, the statistical results for a set of primary comparisons are largely consistent despite substantial changes in the number of secondary comparisons included within the family (22).

Table 6. *Calculations for the false discovery rate procedure*

Comparison i: H_0	P_i	d_i^*
16: $\mu_3 = \mu_1$	0.723	0.050
15: $\mu_8 = \mu_4$	0.369	0.047
14: $\mu_8 = \mu_2$	0.034	0.044
13: $\mu_6 = \mu_4$	0.009	0.041
12: $\mu_7 = \mu_1$	0.008	0.038
11: $\mu_{10} = \mu_6$	0.004	0.034
10: $\mu_4 = \mu_2$	0.003	0.031
9: $\mu_7 = \mu_3$	0.003	0.028
8: $\mu_9 = \mu_5$	0.001^-	0.025
7: $\mu_5 = \mu_3$	0.001^-	0.022
6: $\mu_9 = \mu_1$	0.001^-	0.019
5: $\mu_5 = \mu_1$	0.001^-	0.016
4: $\mu_9 = \mu_7$	0.001^-	0.012
3: $\mu_{10} = \mu_2$	0.001^-	0.009
2: $\mu_6 = \mu_2$	0.001^-	0.006
1: $\mu_{10} = \mu_8$	0.001^-	0.003

P_i, achieved significance level; d_i^*, critical significance level (*Eq. 9*). For P_i, a value of 0.001^- denotes $P_i < 0.001$. If $P_i \leq d_i^*$, then the remaining i null hypotheses are rejected. Because $P_{14} = 0.034 \leq d_{14}^* = 0.044$, null hypotheses 14 $\rightarrow$ 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

[9] Because of the artificial nature of null hypotheses (5), this is a rare occurrence.

[10] If $i < k$ when $P_i \leq d_i^*$, then there will be $k - i$ null hypotheses that cannot be rejected.

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).[11] 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 makeup (9). Therefore, the true error variability is underestimated, and the observed values for the standard deviations of the group means underestimate the true variabilities (9).

To appreciate the impact of correlation on variability, imagine an investigation in a sample of n subjects. In each subject, some random variable X is measured during two experimental conditions: a control period and a subsequent intervention period. Let the random variable measured during the control period be designated X_1 and that during the intervention period be designated X_2. Assume that X_1 and X_2 are distributed normally

$$X_1 \sim N(\mu_1, \sigma_1^2) \quad \text{and} \quad X_2 \sim N(\mu_2, \sigma_2^2)$$

If the random variables 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, $\sigma_{2|1}$, the standard deviation of the conditional distribution of X_2 given that X_1 equals a specific value, depends on the correlation ρ between X_1 and X_2

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

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

1. **Altman DG.** *Practical Statistics for Medical Research.* New York: Chapman & Hall, 1991, p. 210–212.
2. **Armitage P and Berry G.** *Statistical Methods in Medical Research* (3rd ed.). Cambridge, MA: Blackwell Scientific Publications, 1994, p. 224–228.
3. **Benjamini Y and Hochberg Y.** Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 57: 289–300, 1995.
4. **Benjamini Y, Hochberg Y, and Kling Y.** False discovery rate controlling procedures for pairwise comparisons. Working Paper 93-02, Department of Statistics and Operation Research, Tel Aviv University, Tel Aviv, Israel, 1993.
5. **Curran-Everett D, Taylor S, and Kafadar K.** Fundamental concepts in statistics: elucidation and illustration. *J Appl Physiol* 85: 775–786, 1998.
6. **Godfrey K.** Comparing the means of several groups. *N Engl J Med* 313: 1450–1456, 1985.
7. **Hochberg Y and Benjamini Y.** More powerful procedures for multiple significance testing. *Stat Med* 9: 811–818, 1990.
8. **Hsu JC.** *Multiple Comparisons: Theory and Methods.* New York: Chapman & Hall, 1996.
9. **Jones RH.** *Longitudinal Data with Serial Correlation: A State-Space Approach.* New York: Chapman & Hall, 1993.
10. **Koch LG, Meredith TA, Fraker TD, Metting PJ, and Britton SL.** Heritability of treadmill running endurance in rats. *Am J Physiol Regulatory Integrative Comp Physiol* 275: R1455–R1460, 1998.
11. **Ludbrook J.** On making multiple comparisons in clinical and experimental pharmacology and physiology. *Clin Exp Pharmacol Physiol* 18: 379–392, 1991.
12. **Miller RG.** *Simultaneous Statistical Inference* (2nd ed.). New York: Springer-Verlag, 1981.
13. **Moses LE.** *Think and Explain with Statistics.* Reading, MA: Addison-Wesley, 1986, p. 199–203.
14. **Pocock SJ, Hughes MD, and Lee RJ.** Statistical problems in the reporting of clinical trials. *N Engl J Med* 317: 426–432, 1987.
15. **Rothman KJ and Greenland S.** *Modern Epidemiology* (2nd ed.). Philadelphia, PA: Lippincott-Raven, 1998, p. 225–229.
16. **Saville DJ.** Multiple comparison procedures: the practical solution. *Am Statistician* 44: 174–180, 1990.
17. **Shaffer JP.** Multiple hypothesis testing. *Annu Rev Psychol* 46: 561–584, 1995.
18. **Snedecor GW and Cochran WG.** *Statistical Methods* (7th ed.). Ames, IA: The Iowa State University Press, 1980.
19. **Tukey JW.** Some thoughts on clinical trials, especially problems of multiplicity. *Science* 198: 679–684, 1977.
20. **Tukey JW.** The philosophy of multiple comparisons. *Stat Sci* 6: 100–116, 1991.
21. **Tukey JW.** Multiple comparisons: 1948–1983. In: *The Collected Works of John W. Tukey,* edited by Braun HI. New York: Chapman & Hall, 1994, vol. 8.
22. **Williams VSL, Jones LV, and Tukey JW.** Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *J Educ Behav Stat* 24: 42–69, 1999.

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