

HUDM 5123 - Linear Models and Experimental Design

07 - Multiple Comparisons & Type I Error Control

1 Hypothesis Testing

The basic framework for a statistical hypothesis test is to specify a null hypothesis about a population parameter (or parameters) so that, **under the null hypothesis, the probability distribution of some statistic based on the parameter is known**. For example, consider fitting a linear regression model of the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i,$$

under the assumption that $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$. Then

$$t_0 = \frac{\hat{\beta}_k - B}{\text{SE}(\hat{\beta}_k)},$$

has a t -distribution with **$n - p - 1$ degrees** of freedom under the null hypothesis $H_0 : \beta_k = 0$. The p -value for the test is the probability, under the assumption that the null hypothesis is true, of observing a value of t_0 as extreme or more extreme than the one actually observed. **The Type I error rate, α , is the probability that the null hypothesis is incorrectly rejected.** For a given hypothesis test, the convention is to set α to .05, or a one in 20 chance. The logic behind hypothesis testing is based on falsificationism.

1. Assume the null hypothesis is true.
2. Gather data and compute a test statistic.
3. If, under the assumption in (1), the observed test statistic calculated in (2) would have been sufficiently unlikely, reject (1) and conclude the null hypothesis is not true.

The Type I error rate α governs how the term “sufficiently unlikely” is precisely defined. For most researchers in the academic community of social and behavioral scientists, 1/20 is the convention used. How high is too high? Clearly, we (the research community) would not be satisfied if our hypothesis tests had Type I error rates of .25, right? That would mean that even when there is not effect (i.e., the regression coefficient is truly zero) we find an effect one quarter (25%) of the time. Where should we draw the line? It’s arbitrary to some extent, but, nevertheless must be drawn, so we are stuck with .05 so long as there is a compelling reason to use something else. We use this discussion as a motivation for what can go wrong when multiple tests are run with the same data.

2 Multiple Comparisons

The omnibus null hypothesis for a one-way ANOVA where the factor A has k levels is

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k.$$

If the null hypothesis is rejected we may conclude there is **at least one pairwise difference** in means. However, typical research questions require deeper probing into how, precisely, the group means differ from one another. As we have seen, one way to probe more deeply is through examining *pairwise comparisons*. A **pairwise comparison of group means is a hypothesis test of a linear contrast that pits one group mean against another**. In general, for the k -group case, there are $\binom{4}{2} = \frac{4!}{(4-2)!2!} = 6$ distinct pairwise comparisons based on the following linear contrasts.

$$\begin{array}{rclcl}
 \psi_1 = & 1 * \mu_1 + & (-1) * \mu_2 + & 0 * \mu_3 + & 0 * \mu_4 \\
 \psi_2 = & 1 * \mu_1 + & 0 * \mu_2 + & (-1) * \mu_3 + & 0 * \mu_4 \\
 \psi_3 = & 1 * \mu_1 + & 0 * \mu_2 + & 0 * \mu_3 + & (-1) * \mu_4 \\
 \psi_4 = & 0 * \mu_1 + & 1 * \mu_2 + & (-1) * \mu_3 + & 0 * \mu_4 \\
 \psi_5 = & 0 * \mu_1 + & 1 * \mu_2 + & 0 * \mu_3 + & (-1) * \mu_4 \\
 \psi_6 = & 0 * \mu_1 + & 0 * \mu_2 + & 1 * \mu_3 + & (-1) * \mu_4
 \end{array}$$

The null hypothesis for each pairwise comparison is that the contrast is equal to zero. For the first, for example, $H_0 : \psi_1 = 0$ is identical to $H_0 : \mu_1 = \mu_2$. Thus, by testing all six contrasts we are able to determine if there is any evidence for any particular pairs of group means being different. Suppose each of those six tests is tested with a nominal Type I error rate of $\alpha = .05$. Suppose, for the sake of argument, that there are no group differences. **Assuming the tests are statistically independent from one another**, what is the probability that **at least one of the six tests will reject its respective null hypothesis out of the set of six tests**? In other words, we are asking what is the probability that at least one Type I error is committed?

$$\begin{aligned}
 P(\text{At least 1 error}) &= 1 - P(\text{No Type I errors}) \\
 &= 1 - P(\text{No TI err on } \psi_1, \text{No TI err on } \psi_2, \dots, \text{No TI err on } \psi_6) \\
 &= 1 - P(\text{No TI err on } \psi_1) * P(\text{No TI err on } \psi_2) * \dots * P(\text{No TI err on } \psi_6) \\
 &= 1 - .95 * .95 * .95 * .95 * .95 * .95 \\
 &= .265
 \end{aligned}$$

In general, **as the number of comparisons in a set, or family, of comparisons increases, so does the probability of making at least one Type I error**.

$$P(\text{At least one Type I error}) = 1 - (1 - \alpha)^c,$$

where c is the number of tests. Consider a set of m null hypotheses denoted by H_1, H_2, \dots, H_m . Each is tested with a statistical hypothesis test and either rejected if the test is significant or not rejected if the test is not significant. Each test may be classified based on (a) whether the corresponding null hypothesis is true or false and (b) whether the hypothesis is rejected or not as follows.

	H_0 true	H_0 false	Total
H_0 rejected	V	S	R
H_0 not rejected	U	T	$m - R$
Total	m_0	$m - m_0$	m

Table 1: Categorization of hypothesis test results for a set of m tests based on truth of the null hypotheses and test decisions

Here, V represents the number of Type I errors, S the number of correct rejections, and R the total number of rejected null hypotheses.

- The *per-comparison* error rate, α_{PC} , is the probability of Type I error for an individual test (typically .05).
- The *familywise error rate*, α_{FWER} , is the probability of making at least one Type I error for a family of tests. That is, $\alpha_{FWER} = P(V \geq 1) = 1 - P(V = 0)$.
- The *false discovery rate*, or FDR, is expected value of the ratio of false rejections to total rejections: $E[V/R]$.

Methods for controlling familywise error rate are most commonly used in experimental design and analysis. We will consider four strategies for familywise Type I error rate control based on whether the contrasts consist of all pairwise contrasts or not and whether they are planned *a priori* or not. We will consider one method for control of false discovery rate (FDR), which is especially appropriate if the number of tests is very large.

3 Four Methods for Familywise Type I Error Rate Control

As an example, consider a five-group experiment where the factor of interest is the dose in mg of a drug. The five groups are 0 mg (placebo), 10 mg, 20 mg, 30 mg, and 40 mg,. Suppose all the pairwise comparisons are of interest. There are $\binom{5}{2} = 5!/(2!3!) = 10$ distinct pairwise comparisons. Suppose the p-values for the ten comparisons are as follows:

- 0 mg vs 10 mg: $t(256) = .48$; $p = .63$
- 0 mg vs 20 mg: $t(256) = 1.10$; $p = .27$
- 0 mg vs 30 mg: $t(256) = 2.77$; $p = .006$
- 0 mg vs 40 mg: $t(256) = 3.95$; $p = .0001$
- 10 mg vs 20 mg: $t(256) = .58$; $p = .56$
- 10 mg vs 30 mg: $t(256) = 1.28$; $p = .20$
- 10 mg vs 40 mg: $t(256) = 3.12$; $p = .002$
- 20 mg vs 30 mg: $t(256) = .71$; $p = .48$

- 20 mg vs 40 mg: $t(256) = 2.29$; $p = .023$
- 30 mg vs 40 mg: $t(256) = .88$; $p = .38$

Ranking the p-values from smallest to largest gives the following:

.0001, .002, .006, .023, .20, .27, .38, .48, .56, .63

Also suppose that the F statistic for the ANOVA omnibus null hypothesis here has 5 numerator and 256 denominator degrees of freedom.

	Planned – <i>a priori</i>	Not Planned – <i>post hoc</i>
All pairwise	Shaffer's planned post-omnibus	Tukey's HSD
Not all pairwise	Holm-Bonferroni	Scheffé

Table 2: Familywise Type I error rate strategies by scenario

3.1 Post Hoc (Unplanned) Comparisons

Unplanned comparisons are those which are not described in your research questions when you plan a study. You might wish to run unplanned comparisons if you ran a study, gathered data, and the results were not what you expected, in that your original research hypotheses were not confirmed. In an attempt to better understand the data, you might consider exploring for interesting patterns or trends that were not spelled out in your research questions. Without some form of Type I error rate control that accounts for the unplanned nature of the tests, this type of practice can lead to inflated error rates.

The procedure you select for Type I error rate control with unplanned contrast comparisons should depend on **whether the contrasts are pairwise comparisons or not**.

3.1.1 Not All Pairwise - The **Scheffé Correction**

The most **conservative** correction for multiple comparisons we will find reason recommend is the Scheffé correction. It is based on considering the sampling (i.e., probability) distribution of the largest possible F statistic value for **any** contrast in the data. While finding this probability distribution may seem like a daunting task, what Scheffé showed is that for **any** single df contrast that can be constructed, the **sum of squares for that contrast will always be less than or equal to the sum of squares for the associated factor**.

Scheffé then showed that using $(a - 1)F_{.05}(a - 1, N - a)$, where a is the number of groups, **as the F critical value for rejection of the contrast test null hypotheses will hold Type I error rate to the nominal .05 level for an *unlimited* number of unplanned comparisons**.

The first step in calculating the Scheffé-adjusted critical value for testing contrasts is to determine the $F_{.05}(a - 1, N - a)$ critical value. For the example above, $a = 5$ and $N - a$ is 256. $F_{.05}(4, 256)$ may be found using an online calculator, or by using the `qf()` function in R as follows.

```
> qf(p = .95, df1 = 4, df2 = 256)
[1] 2.406905
```

The second step is to multiply the critical value by $a - 1$. Thus, the Scheffé-adjusted F critical value for testing contrasts is $F = 4 * 2.41 = 9.64$. The F distribution with 1 numerator degree of freedom and $N - a$ denominator degrees of freedom is equivalent to the square of a t distribution with $N - a$ degrees of freedom. To test the contrasts, compare the t values to the square root of the Scheffé F critical value, $\sqrt{9.64} = 3.10$. Looking at the p-values, the only one significant after Scheffé adjustment is the 0 mg vs 40 mg comparison.

3.1.2 All Pairwise Comparisons - The Tukey HSD Procedure

It is not uncommon for a research question to center on pairwise differences in group means. Tukey's HSD procedure uses the same logic as the Scheffé correction, with the exception that it capitalizes on the fact that all comparisons are pairwise to create a critical value that is not as extreme as the Scheffé critical value.

Tukey considered the sampling distribution for $F_{\text{pairwise max}}$, the maximum pairwise difference among means, and showed $\sqrt{2F_{\text{pairwise max}}}$ has a "studentized range" distribution under the null hypothesis if assumptions are met. He used the letter q to represent this value as $q = \sqrt{2F_{\text{pairwise max}}}$. In other words, the critical value is $q^2/2$. You will find a table of q values for the studentized range distribution at the end of these notes.

Consulting the table for the example above, first go to the df_{error} column and find the value closest to 256. Here, it is ∞ . Then, go along the top of the table to find the number of means, 5. Finally, select the value that corresponds with the desired familywise Type I error rate; here we will use .05. The value is $q = 3.86$. Thus, the F critical value is $q^2/2 = 3.86^2/2 = 7.45$. Finally, the t critical value is $\sqrt{(7.45)} = 2.73$. Therefore, the 0 mg vs 40 mg, 10 mg vs 40 mg, and 0 mg vs 30 mg comparisons are significant under Tukey's modification.

Several modifications of Tukey's HSD procedure have been proposed for the case when homogeneity of variance is violated. Maxwell & Delaney recommend "Dunnett's T3" if sample sizes are small (i.e., less than 50 per group) and "Games-Howell" if samples sizes are not small.

3.2 A Priori (Planned)

Our two main goals in using a procedure to control the familywise Type I error rate are (a) to have a valid procedure (i.e., one that holds the actual Type I error rate to the nominal level) and (b) given (a), to choose a procedure that provides as much power as possible to detect differences that actually exist. When contrasts are planned in advance alternate methods for familywise Type I error rate control can be used that are more powerful than Scheffé and Tukey's HSD.

3.2.1 Not All Pairwise - The Holm-Bonferroni Correction

The Bonferroni correction may be used for any set of planned comparisons. The procedure is widely used mainly because of its simplicity. For any comparison in a family of comparisons, the Bonferroni correction is implemented by setting α_{PC} equal to $\alpha_{\text{familywise}}$ divided by the number of comparisons being tested.

For example, to hold the familywise error rate to .05 with ten tests in the family, as in the example above, each of the ten planned pairwise comparisons would be tested at $\alpha_{PC}/10 = .05/10 = .005$. Comparing the p-values to .005, we find that the 0 mg vs 40 mg and the 10 mg vs 40 mg comparisons are significant under the Bonferroni procedure

The Holm-Bonferroni procedure is a modification of the Bonferroni procedure that is always at least as powerful as the Bonferroni procedure. The Holm-Bonferroni procedure requires that the p-values be ranked from most to least significant (i.e., smallest p-value to largest p-value). The first p-value is compared with the Bonferroni rejection level of .05/10. If that one is not rejected, we stop and conclude that none are significant; if it is rejected, the next most significant p-value is compared with .05/9. If that one is not rejected, stop and declare only the first is significant; if it is rejected, the next most significant p-value is compared with .05/8. If that one is not rejected, stop and declare only the first two significant; if it is rejected, the last p-value is compared with .05/7. And so on.

For the example above, .0001 < .05/10 (reject); .002 < .05/9 (reject); .006 < .05/8 (reject); .023 > .05/7 (stop). Therefore, the 0 mg vs 40 mg, 10 mg vs 40 mg, and 0 mg vs 30 mg comparisons are significant under the Holm-Bonferroni method. The original paper is by Holm (1979) in the Scandinavian Journal of Statistics.

3.3 All Pairwise - Shaffer's Planned Post-Omnibus

Tukey's HSD procedure is not a bad option for testing all pairwise comparisons, even when they are planned. This is because it is more powerful than a Bonferroni adjustment would be for the set of all pairwise comparisons. However, there is a procedure based on the logical implications of *testing the omnibus test first*, called Shaffer's planned post-omnibus procedure, that can be much more powerful than Tukey's procedure. The caveat with Shaffer's procedure is that with more and more groups it becomes more complicated to implement. We will only cover the case of $a = 3$ and $a = 4$ groups. Beyond that, you should consider using Tukey's HSD even for planned pairwise comparisons.

Shaffer's planned post-omnibus procedure is implemented as follows. First, test the overall omnibus null hypothesis. If it is not significant, stop and conclude there are no significant pairwise contrasts. If it is, then calculate p-values for all pairwise comparisons and order them from most to least significant. Then, use logical implications to test each with a modified α_{PC} , as demonstrated below for the three-group case and the four-group case.

For three groups there will be three pairwise comparisons: AB, AC, and BC. Without any additional information, there are three possible ways that the means could be related:

CASE	TRUE RELATIONSHIP			NUMBER OF POSSIBLE TYPE I ERRORS
1.	XXX			3 choose 2 = $3!/(2!1!) = 3$
2.	XX	X		2 choose 2 = $2!/(2!0!) = 1$
3.	X	X	X	0

Shaffer's procedure capitalizes on the logical implications of the rejected omnibus test. The rejection of the omnibus test implies that there is at least one difference in the pairwise means. This moves us from case 1, where 3 Type I errors are possible, to case 2, where

only 1 Type I error is possible. Order the p-values from most to least significant and test as follows:

1. Compare the most significant to 0.05. If not rejected, stop. If rejected,
2. compare the second most significant to 0.05. If not rejected, stop. If rejected,
3. compare the last p-value to 0.05.

For four groups there will be six pairwise comparisons: AB, AC, AD, BC, BD, and CD. As before, the first step is to consider the omnibus null hypothesis. If not rejected, stop and conclude there are no significant pairwise mean differences. If rejected, proceed as detailed below.

If the omnibus null hypothesis is rejected, it tells us that there is at least one pairwise difference in means. Without any additional information, there are five possible ways that the means could be related:

CASE	TRUE RELATIONSHIP				NUMBER OF POSSIBLE TYPE I ERRORS
1.	XXXX				4 choose 2 = $4!/(2!2!) = 6$
2.	XXX		X		3 choose 2 = $3!/(2!1!) = 3$
3.	XX		XX		2 choose 2 + 2 choose 2 = 1 + 1 = 2
4.	XX	X		X	2 choose 2 = 1
5.	X	X	X	X	0

The rejection of the omnibus test implies that there is at least one difference in the pairwise means. This moves us from the XXXX case where 6 Type I errors are possible to the next case where only 3 are possible. We order the p-values from most to least significant and test as follows:

1. Compare the most significant to 0.05/3. If not rejected, stop. If rejected,
2. compare the second most significant to 0.05/3. If not rejected, stop. If rejected,
3. compare the third most significant to 0.05/3. If not rejected, stop. If rejected,
4. compare the fourth most significant to 0.05/3. If not rejected, stop. If rejected,
5. compare the fifth most significant to 0.05/2. If not rejected, stop. If rejected,
6. compare the last p-value to 0.05.

For five groups there will be ten pairwise comparisons: AB, AC, AD, AE, BC, BD, BE, CD, CE and DE. As before, the first step is to consider the omnibus null hypothesis. If not rejected, stop and conclude there are no significant pairwise mean differences. If rejected, proceed as detailed below.

If the omnibus null hypothesis is rejected, it tells us that there is at least one pairwise difference in means. Without any additional information, there are five possible ways that the means could be related:

CASE	TRUE RELATIONSHIP				NUMBER OF POSSIBLE TYPE I ERRORS
1.	XXXXX				5 choose 2 = $5!/(3!2!) = 10$

2.	XXXX		X		4 choose 2 = $4!/(2!2!) = 6$
3.	XXX		XX		3 choose 2 + 2 choose 2 = $3 + 1 = 4$
4.	XX	X	XX		2 choose 2 + 2 choose 2 = $1 + 1 = 2$
5.	XX	X	X	X	2 choose 2 = 1
6.	X	X	X	X	0

The rejection of the omnibus test implies that there is at least one difference in the pairwise means. This moves us from the XXXXX case, where 10 Type I errors are possible, to the next case where only 6 are possible. After a significant omnibus test, we order the p-values from most to least significant and test as follows:

1. Compare the most significant to $0.05/6$. If not rejected, stop. If rejected,
2. compare the second most significant to $0.05/6$. If not rejected, stop. If rejected,
3. compare the third most significant to $0.05/6$. If not rejected, stop. If rejected,
4. compare the fourth most significant to $0.05/6$. If not rejected, stop. If rejected,
5. compare the fifth most significant to $0.05/6$. If not rejected, stop. If rejected,
6. compare the sixth most significant to $0.05/6$. If not rejected, stop. If rejected,
7. compare the seventh most significant to $0.05/6$. If not rejected, stop. If rejected,
8. compare the eighth most significant to $0.05/4$. If not rejected, stop. If rejected,
9. compare the ninth most significant to $0.05/2$. If not rejected, stop. If rejected,
10. compare the last p-value to 0.05.

For the example above, $.0001 < .05/6$ (reject); $.002 < .05/6$ (reject); $.006 < .05/6$ (reject); $.023 > .05/7$ (stop). Therefore, the 0 mg vs 40 mg, 10 mg vs 40 mg, and 0 mg vs 30 mg comparisons are significant under Shaffer's method. A reference for this method is Shaffer's (1986) JASA paper.

4 False Discovery Rate

When making many comparisons, the strategies we have discussed so far become prohibitively conservative. The cost of familywise error rate control with many tests is that many true effects may be missed because they fail to achieve the adjusted threshold. Imagine a Bonferroni correction based on 1000 comparisons: an individual p-value would have to be 1/1000th the magnitude of the familywise error rate in order to be deemed significant.

One solution to the issue of balancing Type I error rate with power considerations is to use a **different criterion for controlling the error rate**. One alternative is called the *false discovery rate*; Benjamini and Hochberg (1995) is a key reference. Whereas the main idea with familywise error rate control is that a single error is equally problematic, no matter how many tests are in the family, the idea with FDR control is that it is the *proportion* of falsely rejected hypothesis out of the total number of rejections that is important. That is, under FDR control, 2 false rejections (Type I errors) out of 10 total rejections is seen as more problematic than 10 false rejections out of 100 total rejections, even though more Type I errors are made in the latter case. Using the notation given in Table 1, FDR control

methods focus on the ratio V/R by bounding the average value of the ratio such that, for example, $E[V/R] \leq .05$.

In general, my preference is to use FWER control methods for experimental analyses because the number of comparisons are typically small (e.g., 3 to 6 or so) and they provide stronger error rate control. I prefer FDR control methods for exploratory research where many hypotheses (i.e., dozens, hundreds, thousands) will be tested. With more than a couple dozen comparisons in a set, the lack of power from using FWER quickly becomes crippling such that FWER seems like a more reasonable compromise. I should note that reasonable people disagree on which kind of error control to use and which method to use within those categories. If you look at the literature you will find that a lot has been written on correcting for multiple comparisons over the years but, nevertheless, data analyses are complex and nuanced and there is no one-size-fits-all cookbook for intelligent error rate control. Thus, it's good to learn the lay of the land and be familiar with your options.

To implement the Benjamini-Hochberg FDR method with data, rank p-values from largest to smallest and compare each to its rank times the overall FDR (typically .05) divided by the number of comparisons. Thus, for the example above, the largest p-value will be compared to $10 * .05/10$; the second largest will be compared to $9 * .05/10$; the third largest to $8 * .05/10$; and so on. The FDR thresholds are .050, .045, .040, .035, .030, .025, .020, .015, .010, .005. Ranked p-values from largest to smallest and FDR thresholds via B-H method:

.630, .560, .480, .380, .270, .200, .023, .006, .002, .0001
 .050, .045, .040, .035, .030, .025, .020, .015, .010, .005

The 0 mg vs 40 mg, 10 mg vs 40 mg, and 0 mg vs 30 mg comparisons are significant under the B-H FDR method.

As you can see, only the smallest p-value is compared with the Bonferroni threshold; the rest are compared with sequentially less stringent thresholds for significance. Methods that control familywise error rate limit the number of false positives for *all tests in the family* under the assumption that the null hypothesis is true to 5%. Methods that control FDR, on the other hand, does not assume the null hypothesis to be true. Instead, an FDR control method will limit the (average value of the) proportion of of false positives *among rejections* to be 5%. A reference is Thissen, Steinberg, & Kuang (JEBS; 2002).

5 Adjusted p-values

Many programs, instead of giving adjusted alpha cut-offs, will provide adjusted p-values instead. The **emmeans** package we have been using in R is one such example. The cut-off approach is simple in that it provides an adjusted alpha threshold such that the raw p-values may be directly compared to that threshold. For example, for the p-value of .006 above, based on a simple Bonferroni adjustment, we would reject if

$$p \leq .05/10.$$

Based on the Holm-Bonferroni procedure, and because the previous two p-values were rejected, we reject the third if

$$p \leq .05/8.$$

The idea of an adjusted p-value is simply to rescale the p-value so that it is on the scale of the original familywise error rate. Then, we can say we will reject based on Bonferroni if $p_{\text{Bonferroni adjusted}} \leq .05$; or we can reject based on Holm-Bonferroni (H-B) if $p_{\text{H-B adjusted}} \leq .05$. In the two cases above, the adjusted p-values are, respectively, $10p$ and $8p$. With $p = .006$, this gives a Bonferroni-adjusted p-value of .06 and a Holm-Bonferroni-adjusted p-value of .048.

TABLE A.4
CRITICAL VALUES OF STUDENTIZED RANGE DISTRIBUTION

<i>r</i> = number of means (Tukey test) or number of steps between ordered means (Newman-Keuls test)																						
<i>df</i> _{error}	α_{FW}	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	α_{FW}	<i>df</i> _{error}
5	.05	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17	7.32	7.47	7.60	7.72	7.83	7.93	8.03	8.12	8.21	.05	5
	.01	5.70	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48	10.70	10.89	11.08	11.24	11.40	11.55	11.68	11.81	11.93	.01	
6	.05	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65	6.79	6.92	7.03	7.14	7.24	7.34	7.43	7.51	7.59	.05	6
	.01	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30	9.48	9.65	9.81	9.95	10.08	10.21	10.32	10.43	10.54	.01	
7	.05	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30	6.43	6.55	6.66	6.76	6.85	6.94	7.02	7.10	7.17	.05	7
	.01	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55	8.71	8.86	9.00	9.12	9.24	9.35	9.46	9.55	9.65	.01	
8	.05	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.80	6.87	.05	8
	.01	4.75	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86	8.03	8.18	8.31	8.44	8.55	8.66	8.76	8.85	8.94	9.03	.01	
9	.05	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87	5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.64	.05	9
	.01	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49	7.65	7.78	7.91	8.03	8.13	8.23	8.33	8.41	8.49	8.57	.01	
10	.05	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72	5.83	5.93	6.03	6.11	6.19	6.27	6.34	6.40	6.47	.05	10
	.01	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36	7.49	7.60	7.71	7.81	7.91	7.99	8.08	8.15	8.23	.01	
11	.05	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61	5.71	5.81	5.90	5.98	6.06	6.13	6.20	6.27	6.33	.05	11
	.01	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13	7.25	7.36	7.46	7.56	7.65	7.73	7.81	7.88	7.95	.01	
12	.05	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.51	5.61	5.71	5.80	5.88	5.95	6.02	6.09	6.15	6.21	.05	12
	.01	4.32	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94	7.06	7.17	7.26	7.36	7.44	7.52	7.59	7.66	7.73	.01	
13	.05	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43	5.53	5.63	5.71	5.79	5.86	5.93	5.99	6.05	6.11	.05	13
	.01	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79	6.90	7.01	7.10	7.19	7.27	7.35	7.42	7.48	7.55	.01	
14	.05	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46	5.55	5.64	5.71	5.79	5.85	5.91	5.97	6.03	.05	14
	.01	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66	6.77	6.87	6.96	7.05	7.13	7.20	7.27	7.33	7.39	.01	
15	.05	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31	5.40	5.49	5.57	5.65	5.72	5.78	5.85	5.90	5.96	.05	15
	.01	4.17	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55	6.66	6.76	6.84	6.93	7.00	7.07	7.14	7.20	7.26	.01	
16	.05	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26	5.35	5.44	5.52	5.59	5.66	5.73	5.79	5.84	5.90	.05	16
	.01	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46	6.56	6.66	6.74	6.82	6.90	6.97	7.03	7.09	7.15	.01	

<i>r</i> = number of means (Tukey test) or number of steps between ordered means (Newman-Keuls test)																						
<i>df</i> _{error}	α_{FW}	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	α_{FW}	<i>df</i> _{error}
17	.05	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21	5.31	5.39	5.47	5.54	5.61	5.67	5.73	5.79	5.84	.05	17
	.01	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38	6.48	6.57	6.66	6.73	6.81	6.87	6.94	7.00	7.05	.01	
18	.05	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17	5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79	.05	18
	.01	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31	6.41	6.50	6.58	6.65	6.73	6.79	6.85	6.91	6.97	.01	
19	.05	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14	5.23	5.31	5.39	5.46	5.53	5.59	5.65	5.70	5.75	.05	19
	.01	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25	6.34	6.43	6.51	6.58	6.65	6.72	6.78	6.84	6.89	.01	
20	.05	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11	5.20	5.28	5.36	5.43	5.49	5.55	5.61	5.66	5.71	.05	20
	.01	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19	6.28	6.37	6.45	6.52	6.59	6.65	6.71	6.77	6.82	.01	
24	.05	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10	5.18	5.25	5.32	5.38	5.44	5.49	5.55	5.59	.05	24
	.01	3.96	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02	6.11	6.19	6.26	6.33	6.39	6.45	6.51	6.56	6.61	.01	
30	.05	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92	5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.47	.05	30
	.01	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85	5.93	6.01	6.08	6.14	6.20	6.26	6.31	6.36	6.41	.01	
40	.05	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82	4.90	4.98	5.04	5.11	5.16	5.22	5.27	5.31	5.36	.05	40
	.01	3.82	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60	5.69	5.76	5.83	5.90	5.96	6.02	6.07	6.12	6.16	6.21	.01	
60	.05	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81	4.88	4.94	5.00	5.06	5.11	5.15	5.20	5.24	.05	60
	.01	3.76	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45	5.53	5.60	5.67	5.73	5.78	5.84	5.89	5.93	5.97	6.01	.01	
120	.05	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.64	4.71	4.78	4.84	4.90	4.95	5.00	5.04	5.09	5.13	.05	120
	.01	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.37	5.44	5.50	5.56	5.61	5.66	5.71	5.75	5.79	5.83	.01	
∞	.05	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55	4.62	4.68	4.74	4.80	4.85	4.89	4.93	4.97	5.01	.05	∞
	.01	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23	5.29	5.35	5.40	5.45	5.49	5.54	5.57	5.61	5.65	.01	

This table is abridged from Table 29 in E. S. Pearson and H. O. Hartley (Eds.), *Biometrika tables for statisticians* (3rd ed., Vol. 1), Cambridge University Press, New York, 1970, by permission of the *Biometrika* Trustees.