

# Chapter 3

## Extensions

### EXTENSION 3E1: TESTS OF REPLICATION

In Chapter 3, we assumed that the only comparison of interest in the two-group case is that between a cell mean model and a grand mean model. That is, we have compared the full model of

$$Y_{ij} = \mu_j + \varepsilon_{ij}, \quad (3.27, \text{repeated})$$

with the model obtained when we impose the restriction that  $\mu_1 = \mu_2 = \mu$ . However, this is certainly not the only restriction on the means that would be possible. Occasionally, you can make a more specific statement of the results you expect to obtain. This is most often true when your study is replicating previous research that provided detailed information about the phenomena under investigation. As long as you can express your expectation as a restriction on the values of a linear combination of the parameters of the full model, the same general form of our  $F$  test allows you to carry out a comparison of the resulting models.

For example, you may wish to impose a restriction similar to that used in the one-group case in which you specify the exact value of one or both of the population means present in the full model. To extend the numerical example involving the hyperactive-children data presented in Table 3.1 of the text, we might hypothesize that a population of hyperactive children and a population of nonhyperactive children would both have a mean IQ of 98—that is,

$$\mu_1 = \mu_2 = 98. \quad (3E1.1)$$

In this case, our restricted model would simply be

$$Y_{ij} = 98 + \varepsilon_{ij}. \quad (3E1.2)$$

Thus no parameters need to be estimated, and hence the degrees of freedom associated with the model would be  $n_1 + n_2$ .

As a second example, you may wish to specify numerical values for the population means in your restriction but allow them to differ between the two groups. This also would arise in

situations in which you are replicating previous research. Perhaps you carried out an extensive study of hyperactive children in one school year and found the mean IQ of all identified hyperactive children was 106, whereas that of the remaining children was 98. If two years later you wondered whether the values had remained the same and wanted to make a judgment on the basis of a sample of the cases, you could specify these exact values as your null hypothesis or restriction. That is, your restricted model would be

$$\begin{aligned} Y_{i1} &= 106 + \varepsilon_{i1}, \\ Y_{i2} &= 98 + \varepsilon_{i2}. \end{aligned} \quad (3E1.3)$$

Once again, no parameters must be estimated, and so  $df_R = n_1 + n_2$ . As with any model, the sum of squared deviations from the specified parameter values could be used as a measure of the adequacy of this model and be compared with that associated with the full model.

In general, if we let  $c_j$  stand for the constant specified in such a restriction, we could write our restricted model as

$$\begin{aligned} Y_{i1} &= c_1 + \varepsilon_{i1}, \\ Y_{i2} &= c_2 + \varepsilon_{i2}, \end{aligned}$$

or equivalently,

$$Y_{ij} = c_j + \varepsilon_{ij}.$$

The error term used as a measure of the adequacy of such a model would then be

$$E_R = \sum_j \sum_i e_{ijR}^2 = \sum_j \sum_i (Y_{ij} - c_j)^2. \quad (3E1.4)$$

As a third example, you may wish to specify only that the difference between groups is equal to some specified value. Thus if the hyperactive-group mean had been estimated at 106 and the normal-group mean at 98, you might test the hypothesis with a new sample that the hyperactive mean would be 8 points higher than the normal mean. This would allow for the operation of factors such as changing demographic characteristics of the population being sampled, which might cause the IQ scores to generally increase or decrease. The null hypothesis could still be stated easily as  $\mu_1 - \mu_2 = 8$ . It is a bit awkward to state the restricted model in this case, but thinking through the formulation of the model illustrates again the flexibility of the model-comparison approach. In this case, we do not wish to place any constraints on the grand mean, yet we wish to specify the magnitude of the between-group difference at 8 points. We can accomplish this by specifying that the hyperactive-group mean will be 4 points above the grand mean and that the normal-group mean will be 4 points below the grand mean—that is,

$$\begin{aligned} Y_{i1} &= \mu + 4 + \varepsilon_{i1}, \\ Y_{i2} &= \mu - 4 + \varepsilon_{i2}. \end{aligned} \quad (3E1.5)$$

Arriving at a least-squares estimate of  $\mu$  in this context is a slightly different problem than encountered previously. However, we can solve the problem by translating it into a form we

considered in the one-group case. By subtracting 4 from both sides of the equation for the  $Y_{i1}$  scores and adding 4 to both sides of the equation for the  $Y_{i2}$  scores in Equation 3E1.5, we obtain

$$\begin{aligned} Y_{i1} - 4 &= \mu + \varepsilon_{i1}, \\ Y_{i2} + 4 &= \mu + \varepsilon_{i2}. \end{aligned} \quad (3E1.6)$$

This is now essentially the same estimation problem that we used to introduce the least-squares criterion in the one-sample case. There we showed that the least-squares estimate of  $\mu$  is the mean of all scores on the left side of the equations, which here would imply taking the mean of a set of transformed scores, with the scores from Group 1 being 4 less than those observed and the scores in Group 2 being 4 greater than those observed. In the equal- $n$  case, these transformations cancel each other, and the estimate of  $\mu$  would be the same as in a conventional restricted model. In the unequal- $n$  case, the procedure described would generally result in a somewhat different estimate of the grand mean, with the effect that the predictions for the larger group are closer to the mean for that group than is the case for the smaller group. In any event, the errors of prediction are generally different for this restricted model than for a conventional model. In this case, we have

$$E_R = \sum_i (Y_{i1} - 4 - \hat{\mu})^2 = \sum_i (Y_{i2} + 4 - \hat{\mu})^2, \quad (3E1.7)$$

where  $\hat{\mu}$  is the mean of the transformed scores, as described previously.

This test, like the others considered in this chapter, assumes that the population variances of the different groups are equal. We discuss this assumption in more detail in the section of Chapter 3 entitled “Statistical Assumptions” and present procedures there for testing the assumption. In the case in which it is concluded that the variances are heterogeneous, refer to Wilcox (1985) for an alternative procedure for determining if the difference between two-group means differ by more than a specified constant. Additional techniques for imposing constraints on combinations of parameter values are considered in the following chapters.

To try to prevent any misunderstanding that might be suggested by the label *test of replication*, we should stress that the tests we introduce in this section follow the strategy of identifying the constraint on the parameters with the restricted model or the null hypothesis being tested. This allows one to detect if the data depart significantly from what would be expected under this null hypothesis. A significant result then would mean a *failure* to replicate. Note that the identification here of the theoretical expectation with the null hypothesis is different from the usual situation in psychology and instead approximates that in certain physical sciences. As mentioned in Chapter 1, Meehl (1967) calls attention to how theory testing in psychology is usually different from theory testing in physics. In physics, one typically proceeds by making a specific point prediction and assessing whether the data depart significantly from that theoretical prediction, whereas in psychology, one typically lends support to a theoretical hypothesis by rejecting a null hypothesis of no difference. On the one hand, the typical situation in psychology is less precise than that in physics in that the theoretical prediction is often just that “the groups will differ” rather than specifying by how much. On the other hand, the identification of the theoretical prediction with the null hypothesis raises a different set of problems in that the presumption in hypothesis testing is in favor of the null hypothesis. Among the potential disadvantages to such an approach, which applies to the tests of replication introduced here, is that one could be more likely to confirm one’s theoretical expectations by running fewer subjects or doing other things to lower power. It is possible to both have the advantages of a theoretical point prediction and give the presumption to a hypothesis that is different from such theoretical expectations, but doing so requires use of

novel methods beyond what we introduce here. For a provocative discussion of a method of carrying out a test in which the null hypothesis is that data depart by a prespecified amount or more from expectations so that a rejection would mean significant support for a theoretical point prediction, see Serlin and Lapsley (1985). Alternatively, one might use confidence intervals around a mean difference to demonstrate the equivalence of two groups (see discussion in Chapter 4 of Equations 40 and 41; cf. Maxwell, Lau, & Howard, 2015).

## EXTENSION 3E2: ROBUST METHODS FOR ONE-WAY BETWEEN-SUBJECT DESIGNS—BROWN–FORSYTHE, WELCH, AND KRUSKAL–WALLIS TESTS

In Chapter 3, we state that ANOVA is predicated on three assumptions: normality, homogeneity of variance, and independence of observations. When these conditions are met, ANOVA is a “uniformly most powerful” procedure. In essence, this means that the  $F$  test is the best possible test when one is interested uniformly (i.e., equally) in all possible alternatives to the null hypothesis. Thus, in the absence of planned comparisons, ANOVA is the optimal technique to use for hypothesis testing whenever its assumptions hold. In practice, the three assumptions are often met at least closely enough so that the use of ANOVA is still optimal.

Recall from our discussion of statistical assumptions in Chapter 3 that ANOVA is generally robust to violations of normality and homogeneity of variance, although robustness to the latter occurs only with equal  $n$  (more on this later). *Robustness* means that the actual rate of Type I errors committed is close to the nominal rate (typically .05) even when the assumptions fail to hold. In addition, ANOVA procedures generally appear to be robust with respect to Type II errors as well, although less research has been conducted on the Type II error rate.

The general robustness of ANOVA was taken for granted by most behavioral researchers during the 1970s, based on findings documented in the excellent literature review by Glass, Peckham, and Sanders (1972). Because both Type I and Type II error rates were only very slightly affected by violations of normality or homogeneity (with equal  $n$ ), there seemed to be little need to consider alternative methods of hypothesis testing.

However, the 1980s saw a renewed interest in possible alternatives to ANOVA. Although part of the impetus behind this movement stemmed from further investigation of robustness with regard to the Type I error rate, the major focus was on the Type II error rate—that is, on issues of power. As Blair (1981) points out, robustness implies that the power of ANOVA is relatively unaffected by violations of assumptions. However, the user of statistics is interested not in whether ANOVA power is unaffected, but in whether ANOVA is the most powerful test available for a particular problem. Even when ANOVA is robust, it may not provide the most powerful test available when its assumptions have been violated.

Statisticians are developing possible alternatives to ANOVA. Our purpose in this extension is to provide a brief introduction to a few of these possible alternatives. We warn you that our coverage is far from exhaustive; we simply could not cover in a brief review the wide range of possibilities already developed. Instead, our purpose is to make you aware that the field of statistics is dynamic and ever changing, just like other scientific fields of inquiry. Techniques (or theories) that are favored today may be in disfavor tomorrow, replaced by superior alternatives.

Another reason we make no attempt to be exhaustive here is that further research yet needs to be done to compare the techniques we describe for usual ANOVA methods. At this time, it is unclear which, if any, of these methods will be judged most useful. Although we provide evaluative comments where possible, we forewarn you that this area is full of complexity and controversy. The assumption that distributions are normal and variances are homogeneous simplifies

the world enormously. A moment's reflection should convince you that "nonnormal" and "heterogeneous" lack the precision of "normal" and "homogeneous." Data can be nonnormal in an infinite number of ways, rapidly making it very difficult for statisticians to find an optimal technique for analyzing "nonnormal" data. What is good for one form of nonnormality may be bad for another form. Also, what kinds of distributions occur in real data? A theoretical statistician may be interested in comparing data-analysis techniques for data from a specific nonnormal distribution, but if that particular distribution never underlies behavioral data, the comparison may have no practical import to behavioral researchers. How far do actual data depart from normality and homogeneity? There is no simple answer, which partially explains why comparing alternatives to ANOVA is complicated and controversial.

The presentation of methods in this extension is not regularly paralleled by similar extensions on robust methods later in the book because many of the alternatives to ANOVA in the single-factor, between-subjects design have not been generalized to more complex designs.

Two possible types of alternatives to the usual ANOVA in between-subjects designs have received considerable attention in recent years. The first type is a parametric modification of the  $F$  test that does not assume homogeneity of variance. The second type is a nonparametric approach that does not assume normality. Because the third ANOVA assumption is independence, you might expect there to be a third type of alternative that does not assume independence. However, as we stated earlier, independence is largely a matter of design, so modifications would likely involve changes in the design instead of changes in data analysis (see Kenny & Judd, 1986). Besides these two broad types of alternatives, several other possible approaches are being investigated. We look at two of these after we examine the parametric modifications and the nonparametric approaches.

## Parametric Modifications

As stated earlier, one assumption underlying the usual ANOVA  $F$  test is homogeneity of variance. Statisticians have known for many years that the  $F$  test can be either very conservative (too few Type I errors and hence decreased power) or very liberal (too many Type I errors) when variances are heterogeneous and sample sizes are unequal. In general, the  $F$  test is conservative when large sample sizes are paired with large variances. The  $F$  is liberal when large sample sizes are paired with small variances. Extension 3E4, the final extension for Chapter 3, shows why the nature of the pairing causes the  $F$  sometimes to be conservative and other times to be liberal. Obviously, either occurrence is problematic, especially because the population variances are unknown parameters. As a consequence, we can never know with complete certainty whether the assumption has been satisfied in the population. However, statistical tests of the assumption are available (see Chapter 3), so one strategy might be to use the standard  $F$  test to test mean differences only if the homogeneity of variance hypothesis cannot be rejected. Unfortunately, this strategy seems to offer almost no advantage (Wilcox, Charlin, & Thompson, 1986; Zimmerman, 2004). The failure of this strategy has led some statisticians (e.g., Tomarken & Serlin, 1986; Wilcox et al., 1986; Zimmerman, 2004) to recommend that the usual  $F$  test routinely be replaced by one of the more robust alternatives we present here, especially with unequal  $n$ .

Although these problems with unequal  $n$  provide the primary motivation for developing alternatives, several studies have shown that the  $F$  test is not as robust as had previously been thought when sample sizes are equal. Clinch and Keselman (1982), Rogan and Keselman (1977), Tomarken and Serlin (1986), and Wilcox et al. (1986) show that the  $F$  test can become somewhat liberal with equal  $n$  when variances are heterogeneous. When variances are very different from each other, the actual Type I error rate may reach .10 or so (with a nominal rate of .05), even with equal  $n$ . Of course, when variances are less different, the actual error rate is closer to .05.<sup>1</sup> In

summary, there seems to be sufficient motivation for considering alternatives to the  $F$  test when variances are heterogeneous, particularly when sample sizes are unequal.

We consider two alternatives: Brown and Forsythe (1974) developed the first test, which has a rather intuitive rationale. Welch (1951) developed the second test. Both are available in SPSS (one-way ANOVA procedure), so in our discussion, we downplay computational details.<sup>2</sup>

The test statistic developed by Brown and Forsythe (1974) is based on the between-group sum of squares calculated in exactly the same manner as in the usual  $F$  test:

$$SS_B = \sum_{j=1}^a n_j (\bar{Y}_j - \bar{Y})^2, \quad (3E2.1)$$

where  $\bar{Y} = \sum_{j=1}^a n_j \bar{Y}_j / N$ . However, the denominator is calculated differently from the denominator of the usual  $F$  test. The Brown–Forsythe denominator is chosen to have the same expected value as the numerator if the null hypothesis is true, even if variances are heterogeneous. (The rationale for finding a denominator with the same expected value as the numerator if the null hypothesis is true is discussed in Chapter 10.) After some tedious algebra, it can be shown that the expected value of  $SS_B$  under the null hypothesis is given by

$$\mathcal{E}(SS_B) = \sum_{j=1}^a [1 - (n_j / N)] \sigma_j^2. \quad (3E2.2)$$

Notice that if we were willing to assume homogeneity of variance, Equation 3E2.2 would simplify to

$$\mathcal{E}(SS_B) = \sum_{j=1}^a [1 - (n_j / N)] \sigma^2 = \sigma^2 \left[ a - \sum_{j=1}^a (n_j / N) \right] = (a-1)\sigma^2,$$

where  $\sigma^2$  denotes the common variance. With homogeneity,  $\mathcal{E}(MS_W) = \sigma^2$ , so the usual  $F$  is obtained by taking the ratio of  $MS_B$  (which is  $SS_B$  divided by  $a-1$ ) and  $MS_W$ . Under homogeneity,  $MS_B$  and  $MS_W$  have the same expected value under the null hypothesis, so their ratio provides an appropriate test statistic.<sup>3</sup>

When we are unwilling to assume homogeneity, it is preferable to estimate the population variance of each group (i.e.,  $\sigma_j^2$ ) separately. This is easily accomplished by using  $s_j^2$  as an unbiased estimate of  $\sigma_j^2$ . A suitable denominator can be obtained by substituting  $s_j^2$  for  $\sigma_j^2$  in Equation 3E2.2, yielding

$$\sum_{j=1}^a [1 - (n_j / N)] s_j^2. \quad (3E2.3)$$

The expected value of this expression equals the expected value of  $SS_B$  under the null hypothesis, even if homogeneity fails to hold. Thus taking the ratio of  $SS_B$  and the expression in Equation 3E2.3 yields an appropriate test statistic:

$$F^* = \frac{\sum_{j=1}^a n_j (\bar{Y}_j - \bar{Y})^2}{\sum_{j=1}^a [1 - (n_j / N)] s_j^2}. \quad (3E2.4)$$

The statistic is written as  $F^*$  instead of  $F$  because it does not have an exact  $F$  distribution. However, Brown and Forsythe show that the distribution of  $F^*$  can be approximated by an  $F$  distribution

with  $a - 1$  numerator degrees of freedom and  $f$  denominator degrees of freedom. Unfortunately, the denominator degrees of freedom are tedious to calculate and are best left to a computer program. Nevertheless, we present the formula for denominator degrees of freedom as follows:

$$f = \frac{1}{\sum_{j=1}^a \frac{g_j^2}{(n_j - 1)}}, \quad (3E2.5)$$

where

$$g_j = \frac{[1 - (n_j / N)] s_j^2}{\sum_{j=1}^a [1 - (n_j / N)] s_j^2}.$$

It is important to realize that, in general,  $F^*$  differs from  $F$  in two ways. First, the denominator degrees of freedom for the two approaches are different. Second, the observed values of the test statistics are typically different as well. In particular,  $F^*$  may be either systematically smaller or larger than  $F$ . If large samples are paired with small variances,  $F^*$  tends to be smaller than  $F$ ; however, this reflects an advantage for  $F^*$ , because  $F$  tends to be liberal in this situation. Conversely, if large samples are paired with large variances,  $F^*$  tends to be larger than  $F$ ; once again, this reflects an advantage for  $F^*$ , because  $F$  tends to be conservative in this situation. What if sample sizes are equal? With equal  $n$ , Equation 3E2.4 can be rewritten as

$$\begin{aligned} F^* &= \frac{n \sum_{j=1}^a (\bar{Y}_j - \bar{Y})^2}{\sum_{j=1}^a [1 - (1/a)] s_j^2} = \frac{n \sum_{j=1}^a n_j (\bar{Y}_j - \bar{Y})^2}{\sum_{j=1}^a [(a-1)/a] s_j^2} \\ &= \frac{n \sum_{j=1}^a (\bar{Y}_j - \bar{Y})^2}{(a-1) \sum_{j=1}^a s_j^2 / a} = \frac{n \sum_{j=1}^a (\bar{Y}_j - \bar{Y})^2 / (a-1)}{\sum_{j=1}^a s_j^2 / a} \\ &= \frac{MS_B}{MS_W} = F. \end{aligned}$$

Thus, with equal  $n$ , the observed values of  $F^*$  and  $F$  are identical. However, the denominator degrees of freedom are still different. It can be shown that with equal  $n$ , Equation 3E2.5 for the denominator degrees of freedom associated with  $F^*$  becomes

$$f = \frac{(n-1) \left( \sum_{j=1}^a s_j^2 \right)^2}{\sum_{j=1}^a (s_j^2)^2}. \quad (3E2.6)$$

Although it may not be immediately apparent,  $f$  is an index of how different sample variances are from each other. If all sample variances were identical to each other,  $f$  would equal  $a(n - 1)$ ,

the denominator degrees of freedom for the usual  $F$  test. At the other extreme, as one variance becomes infinitely larger than all others,  $f$  approaches a value of  $n - 1$ . In general, then,  $f$  ranges from  $n - 1$  to  $a(n - 1)$  and attains higher values for more similar variances.

We can summarize the relationship between  $F^*$  and  $F$  with equal  $n$  as follows. To the extent that the sample variances are similar,  $F^*$  is similar to  $F$ ; however, when sample variances are different from each other,  $F^*$  is more conservative than  $F$  because the lower denominator degrees of freedom for  $F^*$  imply a higher critical value for  $F^*$  than for  $F$ . As a consequence, with equal  $n$ ,  $F^*$  rejects the null hypothesis less often than does  $F$ . If the homogeneity of variance assumption is valid, the implication is that  $F^*$  is less powerful than  $F$ . However, Monte Carlo studies by Clinch and Keselman (1982) and Tomarken and Serlin (1986) suggest that the power advantage of  $F$  over  $F^*$  rarely exceeds .03 with equal  $n$ .<sup>4</sup> On the other hand, if the homogeneity assumption is violated,  $F^*$  tends to maintain  $\alpha$  at .05, whereas  $F$  becomes somewhat liberal. However, the usual  $F$  test tends to remain robust as long as the population variances are not widely different from each other. As a result, in practice, any advantage that  $F^*$  might offer over  $F$  with equal  $n$  is typically slight, except when variances are extremely discrepant from each other.

However, with unequal  $n$ ,  $F^*$  and  $F$  may be very different from one another. If it so happens that large samples are paired with small variances,  $F^*$  maintains  $\alpha$  near .05 (assuming that .05 is the nominal value), whereas the actual  $\alpha$  level for the  $F$  test can reach .15 or even .20 (Clinch & Keselman, 1982; Tomarken & Serlin, 1986), if population variances are substantially different from each other. Conversely, if large samples happen to be paired with large variances,  $F^*$  provides a more powerful test than does the  $F$  test. The advantage for  $F^*$  can be as great as .15 or .20 (Tomarken & Serlin, 1986), depending on how different the population variances are and on how the variances are related to the sample sizes. Thus  $F^*$  is not necessarily more conservative than  $F$ .

Welch (1951) also derived an alternative to the  $F$  test that does not require the homogeneity of variance assumption. Unlike the Brown and Forsythe alternative, which was based on the between-group sum of squares of the usual  $F$  test, Welch's test uses a different weighting of the sum of squares in the numerator. Welch's statistic is defined as

$$W = \frac{\sum_{j=1}^a w_j (\bar{Y}_j - \bar{Y})^2 / (a-1)}{\left[ 1 + \frac{2}{3} (a-1) \Lambda \right]}$$

where,

$$w_j = n_j / s_j^2$$

$$\tilde{Y} = \frac{\sum_{j=1}^a w_j \bar{Y}_j}{\sum_{j=1}^a w_j}$$

$$\Lambda = \frac{3 \sum_{j=1}^a w_j \left\{ \left[ 1 - w_j / \sum_{j=1}^a w_j \right]^2 \right\}}{a^2 - 1} / (n_j - 1)$$

When the null hypothesis is true,  $W$  is approximately distributed as an  $F$  variable with  $a - 1$  numerator and  $1/\Lambda$  denominator degrees of freedom. (Notice that  $\Lambda$  is used to represent the value of Wilks's lambda in Chapter 14. Its meaning here is entirely different and reflects the unfortunate tradition among statisticians to use the same symbol for different expressions. In any event, the meaning here should be clear from the context.) It might alleviate some concern to remind you at this point that the SPSS program for one-way ANOVA calculates  $W$  as well as its degrees of freedom and associated  $p$  value.

The basic difference between the rationales behind  $F^*$  and  $W$  involves the weight associated with a group's deviation from the grand mean—that is,  $\bar{Y}_j - \bar{Y}$ . As Equation 3E2.1 shows,  $F^*$  weights each group according to its sample size. Larger groups receive more weight because their sample mean is likely to be a better estimate of their population mean.  $W$ , however, weights each group according to  $n_j/s_j^2$ , which is the reciprocal of the estimated variance of the mean. Less variable group means thus receive more weight, whether the lesser variability results from a larger sample size or a smaller variance. This difference in weighting causes  $W$  to be different from  $F^*$ , even though neither assumes homogeneity of variance. As an aside, notice also that the grand mean is defined differently in Welch's approach than for either  $F$  or  $F^*$ ; although it is still a weighted average of the group means, the weights depend on the sample variances as well as the sample sizes.

Welch's  $W$  statistic compares to the usual  $F$  test in a generally similar manner as  $F^*$  compares to  $F$ . When large samples are paired with large variances,  $W$  is less conservative than  $F$ . When large samples are paired with small variances,  $W$  is less liberal than  $F$ . Interestingly, when sample sizes are equal,  $W$  differs more from  $F$  than does  $F^*$ . Whereas  $F$  and  $F^*$  have the same observed value with equal  $n$ , in general, the observed value of  $W$  is different. The reason is that, as seen earlier,  $W$  gives more weight to groups with smaller sample variances. When homogeneity holds in the population, this differential weighting is simply based on chance, because in this situation, sample variances differ from one another as a result of sampling error only. As a result, tests based on  $W$  are somewhat less powerful than tests based on  $F$ . Based on Tomarken and Serlin's (1986) findings, the difference in power is usually .03 or less and would rarely exceed .06 unless sample sizes are very small. However, when homogeneity fails to hold,  $W$  can be appreciably more powerful than the usual  $F$  test, even with equal  $n$ . The power advantage of  $W$  was often as large as .10 and even reached .34 in one condition in Tomarken and Serlin's simulations. This advantage stems from  $W$  giving more weight to the more stable sample means, which  $F$  does not do (nor does  $F^*$ ). It must be added, however, that  $W$  can also have less power than  $F$  with equal  $n$ . If the group that differs most from the grand mean has a large population variance,  $W$  attaches a relatively small weight to the group because of its large variance. In this particular case,  $W$  tends to be less powerful than  $F$  because the most discrepant group receives the least weight. Nevertheless, Tomarken and Serlin found that  $W$  is generally more powerful than  $F$  for most patterns of means when heterogeneity occurs with equal  $n$ .

The choice between  $F^*$  and  $W$  when heterogeneity is suspected is difficult given the current state of knowledge. On the one hand, Tomarken and Serlin (1986) found that  $W$  is more powerful than  $F^*$  across most configurations of population means. On the other hand, Clinch and Keselman (1982) found that  $W$  becomes somewhat liberal when underlying population distributions are skewed instead of normal. They found that  $F^*$  generally maintains a close to nominal value of .05 even for skewed distributions. In addition, Wilcox et al. (1986) found that  $W$  maintained an appropriate Type I error rate better than  $F^*$  when sample sizes are equal, but that  $F^*$  was better than  $W$  when unequal sample

sizes are paired with equal variances. Choosing between  $F^*$  and  $W$  is obviously far from clear-cut, given the complex nature of findings. Further research is needed to clarify their relative strengths. Although the choice between  $F^*$  and  $W$  is unsettled, it is clear that both are preferable to  $F$  when population variances are heterogeneous and sample sizes are unequal.

Table 3E2.1 summarizes the properties of  $F$ ,  $F^*$ , and  $W$  as a function of population variances and sample sizes. Again, from a practical standpoint, the primary point of the table is that  $F^*$  or  $W$  should be considered seriously as a replacement for the usual  $F$  test when sample sizes are unequal and heterogeneity of variance is suspected.

TABLE 3E2.1  
PROPERTIES OF  $F$ ,  $F^*$ , AND  $W$  AS A FUNCTION OF SAMPLE SIZES AND  
POPULATION VARIANCES

	<i>Test Statistic</i>		
	$F$	$F^*$	$W$
<i>Equal Sample Sizes</i>			
Equal variances	Appropriate	Slightly conservative	Robust
Unequal variances	Robust, except can become liberal for very large differences in variances	Robust, except can become liberal for extremely large differences in variances	Robust
<i>Unequal Sample Sizes</i>			
Equal variances	Appropriate	Robust	Robust, except can become slightly liberal for very large differences in sample sizes
Large samples paired with large variances	Conservative	Robust, except can become slightly liberal when differences in sample sizes and in variances are both very large	Robust, except can become slightly liberal when differences in sample sizes and in variances are both very large
Large samples paired with small variances	Liberal	Robust, except can become slightly liberal when differences in sample sizes and in variances are both very large	Robust, except can become slightly liberal when differences in sample sizes and in variances are both very large

## Nonparametric Approaches

The parametric modifications of the previous section were developed for analyzing data with unequal population variances. The nonparametric approaches of this section were developed for analyzing data whose population distributions are nonnormal. As we discuss in some detail later, another motivating factor for the development of nonparametric techniques in the behavioral sciences has been the belief held by some researchers that they require less stringent measurement properties of the dependent variable. The organizational structure of this section consists of, first, presenting a particular nonparametric technique, and, second, discussing its merits relative to parametric techniques.

There are several nonparametric alternatives to ANOVA for the single-factor, between-subjects design. We present only one of these, the Kruskal–Wallis test, which is the most frequently used nonparametric test for this design. For information on other nonparametric methods, consult such nonparametric textbooks as Bradley (1968), Cliff (1996), Gibbons (1971), Marascuilo and McSweeney (1977), Noether (1976), and Siegel (1956).

The Kruskal–Wallis test is often called an “ANOVA by Ranks” because a fundamental distinction between the usual ANOVA and the Kruskal–Wallis test is that the original scores are replaced by their ranks in the Kruskal–Wallis test. Specifically, the first step in the test is to rank order all observations from low to high (actually, high to low yields exactly the same result) in the entire set of  $N$  subjects. Be certain to notice that this ranking is performed across all  $a$  groups, independently of group membership. When scores are tied, each observation is assigned the average (i.e., mean) rank of the scores in the tied set. For example, if three scores are tied for 6th, 7th, and 8th place in order, all three scores are assigned a rank of 7.

Once the scores have been ranked, the test statistic is given by

$$H = \frac{12}{N(N+1)} \sum_{j=1}^a n_j \left\{ \bar{R}_j - [(N+1)/2] \right\}^2, \quad (3E2.7)$$

where  $\bar{R}_j$  is the mean rank for group  $j$ . Although Equation 3E2.7 may look very different from the usual ANOVA  $F$  statistic, in fact, there is an underlying similarity. For example,  $(N+1)/2$  is simply the grand mean of the ranks, which we know must have values of 1, 2, 3, ...,  $N$ . Thus the term  $\sum_{j=1}^a n_j \left\{ \bar{R}_j - [(N+1)/2] \right\}^2$  is a weighted sum of squared deviations of group means from the grand mean, as in the parametric  $F$  test. It also proves to be unnecessary to estimate  $\sigma^2$ , the population error variance, because the test statistic is based on a finite population of size  $N$  (cf. Marascuilo & McSweeney, 1977, for more on this point). The important point for our purposes is that the Kruskal–Wallis test is very much like an ANOVA on ranks.

When the null hypothesis is true,  $H$  is approximately distributed as a  $\chi^2$  with  $a - 1$  degrees of freedom. The  $\chi^2$  approximation is accurate unless sample sizes within some groups are quite small, in which case tables of the exact distribution of  $H$  should be consulted in such sources as Siegel (1956) or Iman, Quade, and Alexander (1975). When ties occur in the data, a correction factor  $T$  should be applied:

$$T = 1 - \frac{\sum_{i=1}^G (t_i^3 - t_i)}{N^3 - N},$$

where  $t_i$  is the number of observations tied at a particular value and  $G$  is the number of distinct values for which there are ties. A corrected test statistic  $H'$  is obtained by dividing  $H$  by  $T$ :  $H' = H/T$ . The correction has little effect (i.e.,  $H'$  differs very little from  $H$ ) unless sample sizes are very small or there are many ties in the data relative to sample size. Most major statistical packages (e.g., SAS and SPSS) have a program for computing  $H$  (or  $H'$ ) and its associated  $p$  value. Also, it should be pointed out that when there are only two groups to be compared (i.e.,  $a = 2$ ), the Kruskal–Wallis test is equivalent to the Wilcoxon Rank Sum test, which is also equivalent to the Mann–Whitney  $U$ .

## Choosing Between Parametric and Nonparametric Tests

Statisticians have debated the relative merits of parametric versus nonparametric tests ever since the inception of nonparametric approaches. As a consequence, all too often behavioral researchers are told either that parametric procedures should always be used (because they are robust and more powerful) or that nonparametric methods should always be used (because they make fewer assumptions). Not surprisingly, both of these extreme positions are oversimplifications. We provide a brief overview of the advantages each approach possesses in certain situations. Our discussion is limited to a comparison of the  $F$ ,  $F^*$ , and  $W$  parametric tests and the Kruskal–Wallis nonparametric test.

Nevertheless, even with this limitation, do not expect our comparison of the methods to provide a definitive answer as to which approach is “best.” The choice between approaches is too complicated for such a simple answer. There are certain occasions where parametric tests are preferable, and there are others where nonparametric tests are better. A wise data analyst carefully weighs the advantages in his or her situation and makes an informed choice accordingly.

A primary reason the comparison of parametric and nonparametric approaches is so difficult is that they do not always test the same null hypothesis. To see why they do not, we must consider the assumptions associated with each approach. As stated earlier, we consider specifically the  $F$  test and Kruskal–Wallis test for one-way between-subjects designs.

As discussed in Chapter 3, the parametric ANOVA can be conceptualized in terms of a full model of the form

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}.$$

ANOVA tests a null hypothesis

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_a = 0,$$

where it is assumed that population distributions are normal and have equal variances. In other words, under the null hypothesis, all  $a$  population distributions are identical normal distributions if ANOVA assumptions hold. If the null hypothesis is false, one or more distributions are shifted either to the left or to the right of the other distributions. Figure 3E2.1 illustrates such an occurrence for the case of three groups. The three distributions are identical except that  $\mu_1 = 10$ ,  $\mu_2 = 20$ , and  $\mu_3 = 35$ . When the normality and homogeneity assumptions are met, the distributions still have the same shape, but they have different *locations* when the null hypothesis is false. For this reason, ANOVA is sometimes referred to as a *test of location* or as *testing a shift hypothesis*.

Under certain conditions, the Kruskal–Wallis test can also be conceptualized as testing a shift hypothesis. However, although it may seem surprising given it has been many years since Kruskal and Wallis (1952) introduced their test, there has been a fair amount of confusion and variation in textbook descriptions even decades later about what assumptions are required and what hypothesis is tested by the Kruskal–Wallis test (a summary is provided by Vargha & Delaney, 1998). As usual, what one can conclude is driven largely by what one is willing to assume. If one is willing to assume that the distributions being compared are identical except possibly for their location, then the Kruskal–Wallis test can lead to a similar conclusion as an ANOVA. Like other authors who adopt these restrictive “shift model”

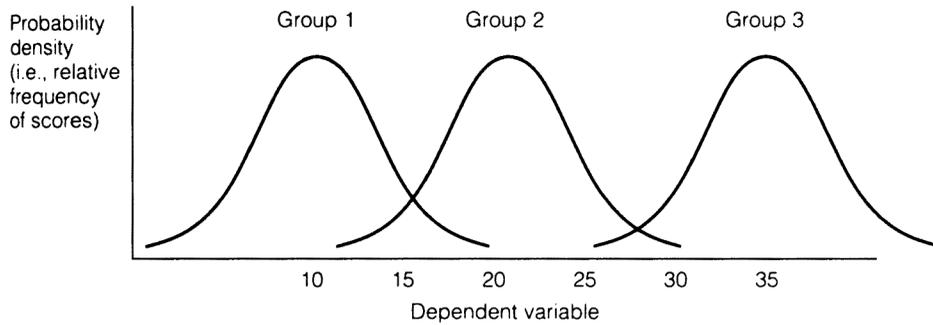


FIGURE 3E2.1 Shifted distributions under ANOVA assumptions

assumptions, Hollander and Wolfe (1973) argue that the Kruskal–Wallis test can be thought of in terms of a full model of the form

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

and that the null hypothesis being tested can be represented by

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_a = 0,$$

just as in the parametric ANOVA. From this perspective, the only difference concerns the assumptions involving the distribution of errors ( $\varepsilon_{ij}$ ). Whereas the parametric ANOVA assumes both normality and homogeneity, the Kruskal–Wallis test assumes only that the population of error scores has an identical continuous distribution for every group. As a consequence, in the Kruskal–Wallis model, homogeneity of variance is still assumed, but normality is not. The important point for our purposes is that, under these assumptions, the Kruskal–Wallis test is testing a shift hypothesis, as is the parametric ANOVA, when its assumptions are met. Figure 3E2.2 illustrates such an occurrence for the case of three groups. As in Figure 3E2.1, the three distributions of Figure 3E2.2 are identical to each other except for their location on the  $X$  axis. Notice, however, that the distributions in Figure 3E2.2 are skewed, unlike the

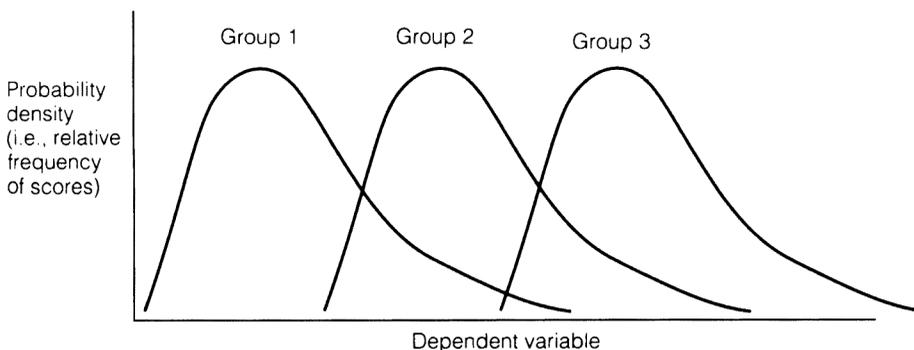


FIGURE 3E2.2 Shifted distributions under Kruskal–Wallis assumptions

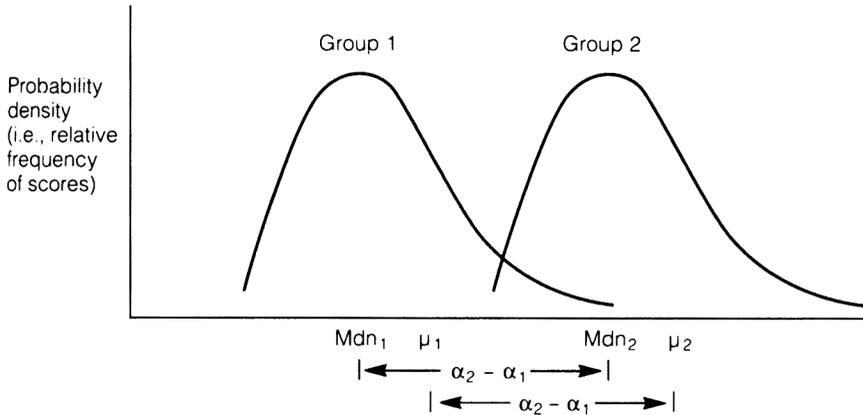


FIGURE 3E2.3 Meaning of  $\alpha_2 - \alpha_1$  for two groups when shift hypothesis holds

distributions in Figure 3E2.1 that are required to be normal by the ANOVA model. Under these conditions, both approaches are testing the same null hypothesis, because the  $\alpha_j$  parameters in the models are identical. For example, the difference  $\alpha_2 - \alpha_1$  represents the extent to which the distribution of Group 1 is shifted either to the right or to the left of Group 2. Not only does  $\alpha_2 - \alpha_1$  equal the difference between the population means, but as Figure 3E2.3 shows, it also equals the difference between the medians, the 5th percentile, the 75th percentile, or any other percentile. Indeed, it is fairly common to regard the Kruskal–Wallis test as a way of deciding if the population medians differ (cf. Wilcox, 1996, p. 365). This is legitimate when the assumption that all distributions have the same shape is met. In this situation, the only difference between the two approaches is that the parametric ANOVA makes the additional assumption that this common shape is that of a normal distribution. Of course, this difference implies different properties for the tests, which we discuss momentarily. To summarize, when one adopts the assumptions of the shift model (namely, identical distributions for all  $a$  groups, except for a possible shift under the alternative hypothesis), the Kruskal–Wallis test and the parametric ANOVA are testing the same null hypothesis. In this circumstance, it is possible to compare the two approaches and state conditions under which each approach is advantageous.<sup>5</sup>

However, there are good reasons for viewing the Kruskal–Wallis test differently. Although it is widely understood that the null hypothesis being tested is that the groups have identical population distributions, the most appropriate alternative hypothesis and required assumptions are not widely understood. One important point to understand is that the Kruskal–Wallis test possesses the desirable mathematical property of consistency only with respect to an alternative hypothesis stated in terms of whether individual scores are greater in one group than another.

This is seen most clearly when two groups are being compared. Let  $p$  be defined as the probability that a randomly sampled observation from one group on a continuous dependent variable  $Y$  is greater than a randomly sampled observation from the second group (e.g., Wilcox, 1996, p. 365). That is,

$$p = \Pr(Y_1 > Y_2).$$

If the two population distributions are identical,  $p = 1/2$ , and so one could state the null hypothesis the Kruskal–Wallis is testing in this case as  $H_0 : p = 1/2$ . The mathematical reason doing so makes most sense is that the test is consistent against an alternative hypothesis if and only if it implies  $H_1 : p \neq 1/2$  (Kendall & Stuart, 1973, p. 513). Cases in which  $p = 1/2$  have been termed cases of *stochastic equality* (cf. Mann & Whitney, 1947; Delaney & Vargha, 2002) and the consistency property means that if the two populations are stochastically unequal ( $p \neq 1/2$ ), then the probability of rejecting the null hypothesis approaches 1 as the sample sizes get larger.

When the populations have the same shape, ANOVA and Kruskal–Wallis are testing the same hypothesis regarding location, although it is common to regard the ANOVA as a test of differences between population means, but regard the Kruskal–Wallis test as a test of differences between population medians.<sup>6</sup> However, when distributions have different asymmetric shapes, it is possible for the population means to all be equal and yet the population medians all be different, or vice versa. Similarly, with regard to the hypothesis that the Kruskal–Wallis test is most appropriate for, namely, stochastic equality with different asymmetric distributions, the population means might all be equal, yet the distributions be stochastically unequal, or vice versa. The point is that, in such a case, the parametric ANOVA may be testing a true null hypothesis, whereas the nonparametric approach is testing a false null hypothesis. In such a circumstance, the probabilities of rejecting the null hypothesis for the two approaches cannot be compared meaningfully because they are answering different questions.

In summary, when distributions have different shapes, the parametric and nonparametric approaches are generally testing different hypotheses. Different shapes occur fairly often, in part because of floor-and-ceiling effects such as occur with Likert-scale dependent variables. In such conditions, the basis for choosing between the approaches should probably involve consideration of whether the research question is best formulated in terms of population means or in terms of comparisons of individual scores. When these differ, we would argue that more often than not, the scientist is interested in the comparison of individuals. If one is interested in comparing two methods of therapy for reducing depression, or clients' daily alcohol consumption, one is likely more interested in which method would help the greater number of people rather than which method produces the greater mean level of change—if these are different. In such situations, stochastic comparison might be preferred to the comparison of means.

Suppose that, in fact, population distributions are identical—which approach is better, parametric or nonparametric? Although the question seems relatively straightforward, the answer is not. Under some conditions, such as normal distributions, the parametric approach is better. However, under other conditions, such as certain long-tailed distributions (in which extreme scores are more likely than in the normal distribution), the nonparametric approach is better. As usual, the choice involves a consideration of Type I error rate and power.

If population distributions are identical and normal, both the  $F$  test and the Kruskal–Wallis test maintain the actual  $\alpha$  level at the nominal value, because the assumptions of both tests have been met (assuming, in addition, as we do throughout this discussion, that observations are independent of one another). On the other hand, if distributions are identical but nonnormal, only the assumptions of the Kruskal–Wallis test are met. Nevertheless, the extensive survey conducted by Glass and colleagues (1972) suggests that the  $F$  test is robust with respect to Type I errors to all but extreme violations of normality.<sup>7</sup> Thus, with regard to Type I error rates, there is little practical reason to prefer either test over the other if all population distributions have identical shapes.

While on the topic of the Type I error rate, it is important to dispel a myth concerning nonparametric tests. Many researchers apparently believe that the Kruskal–Wallis test should be used instead of the  $F$  test when variances are unequal, because the Kruskal–Wallis test does not assume homogeneity of variance. However, we can see that this belief is misguided. Under the

shift model, the Kruskal–Wallis test assumes that population distributions are identical under the null hypothesis, and identical distributions obviously have equal variances. Even when the Kruskal–Wallis is treated as a test of stochastic equality, the test assumes that the *ranks* of the scores are equally variable across groups (Vargha & Delaney, 1998), so homogeneity of variance in some form is, in fact, an assumption of the Kruskal–Wallis test. Furthermore, the Kruskal–Wallis test is not robust to violations of this assumption with unequal  $n$ . Keselman, Rogan, and Feir-Walsh (1977), as well as Tomarken and Serlin (1986), found that the actual Type I error rate of the Kruskal–Wallis test could be as large as twice the nominal level when large samples are paired with small variances (cf., Oshima & Algina, 1992). It should be added that the usual  $F$  test was even less robust than the Kruskal–Wallis test. However, the important practical point is that neither test is robust. In contrast, Tomarken and Serlin (1986) found both  $F^*$  and  $W$  to maintain acceptable  $\alpha$  levels even for various patterns of unequal sample sizes and unequal variances.<sup>8</sup> Thus the practical implication is that  $F^*$  and  $W$  are better alternatives to the usual  $F$  test than is the standard Kruskal–Wallis test when heterogeneity of variance is suspected, especially with unequal  $n$ . Robust forms of the Kruskal–Wallis test have now been proposed, but are not considered here (see Delaney & Vargha, 2002).

A second common myth surrounding nonparametric tests is that they are always less powerful than parametric tests. It is true that if the population distributions for all  $a$  groups are normal with equal variances, then the  $F$  test is more powerful than the Kruskal–Wallis test. The size of the difference in power varies as a function of the sample sizes and the means, so it is impossible to state a single number to represent how much more powerful the  $F$  test is. However, it is possible to determine mathematically that as sample sizes increase toward infinity, the efficiency of the Kruskal–Wallis test to the  $F$  test is .955 under normality.<sup>9</sup> In practical terms, this means that for large samples, the  $F$  test can achieve the same power as the Kruskal–Wallis test and yet require only 95.5% as many subjects as would the Kruskal–Wallis test. It can also be shown that for large samples, the Kruskal–Wallis test is at least 86.4% as efficient as the  $F$  test for distributions of any shape, as long as all  $a$  distributions have the same shape. Thus, at its absolute worst, for large samples, using the Kruskal–Wallis instead of the  $F$  test is analogous to failing to use 13.6% of the subjects one has observed. We must add, however, that the previous statement assumes that all population distributions are identical. If they are not, the Kruskal–Wallis test in some circumstances has little or no power for detecting true mean differences, because it is testing a different hypothesis—namely, stochastic equality.

So far, we have done little to dispel the myth that parametric tests are always more powerful than nonparametric tests. However, for certain nonnormal distributions, the Kruskal–Wallis test is, in fact, considerably more powerful than the parametric  $F$  test. Generally speaking, the Kruskal–Wallis test is more powerful than the  $F$  test when the underlying population distributions are symmetric but heavy-tailed, which means that extreme scores (i.e., outliers) are more frequent than in the normal distribution. The size of the power advantage of the Kruskal–Wallis test depends on the particular shape of the nonnormal distribution, sample sizes, and magnitude of separation between the groups. However, the size of this advantage can easily be large enough to be of practical importance in some situations. It should also be added that the Kruskal–Wallis test is frequently more powerful than the  $F$  test when distributions are identical but skewed.

As mentioned earlier, another argument that has been made for using nonparametric procedures is that they require less stringent measurement properties of the data. In fact, there has been a heated controversy ever since Stevens (1946, 1951) introduced the concept of “levels of measurement” (i.e., nominal, ordinal, interval, and ratio scales) with his views of their implications for statistics. Stevens argues that the use of parametric statistics requires that the observed dependent variable be measured on an interval or ratio scale. However, many behavioral variables fail to meet this criterion, which has been taken by some psychologists to imply that most behavioral

data should be analyzed with nonparametric techniques. Others (e.g., Gaito, 1980; Lord, 1953) argue that the use of parametric procedures is entirely appropriate for behavioral data.

We cannot possibly do justice in this discussion to the complexities of all viewpoints. Instead, we attempt to describe briefly a few themes and recommend additional reading. Gardner (1975) provides an excellent review of both sides of the controversy through the mid-1970s. Three points raised in his review deserve special mention here. First, parametric statistical tests do not make any statistical assumptions about the level of measurement. As we stated previously, the assumptions of the  $F$  test are normality, homogeneity of variance, and independence of observations. A correct numerical statement concerning population mean differences does not require interval measurement. Second, although a parametric test can be performed on ordinal data without violating any assumptions of the test, the meaning of the test could be damaged. In essence, this can be thought of as a potential construct validity problem (see Chapter 2). Although the test is correct as a statement of mean group differences on the observed variable, these differences might not reflect true differences on the underlying construct. Third, Gardner cites two empirical studies (Baker, Hardyck, & Petrinovich, 1966; Labovitz, 1967) that showed that, although in theory construct validity might be problematic, in reality, parametric tests produced meaningful results for constructs even when the level of measurement was only ordinal.

Recent work demonstrates that the earlier empirical studies conducted prior to 1980 were correct as far as they went, but it has become clear that these earlier studies were limited in an important way. In effect, the earlier studies assumed that the underlying population distributions on the construct not only had the same mean but also were literally identical to each other. However, a number of later studies (e.g., Maxwell & Delaney, 1985; Spencer, 1983) show that when the population distributions on the construct have the same mean but different variances, parametric techniques on ordinal data can result in very misleading conclusions. Thus, in some practical situations, nonparametric techniques may indeed be more appropriate than parametric approaches. Many interesting articles continue to be written on this topic. Articles deserving attention are Davison and Sharma (1988), Marcus-Roberts and Roberts (1987), Michell (1986), and Townsend and Ashby (1984).

In summary, the choice between a parametric test ( $F$ ,  $F^*$ , or  $W$ ) and the Kruskal–Wallis test involves consideration of a number of factors. First, the Kruskal–Wallis test does not always test the same hypothesis as the parametric tests. As a result, in general, it is important to consider whether the research question of interest is most appropriately formulated in terms of comparisons of individual scores or comparisons of means. Second, neither the usual  $F$  test nor the Kruskal–Wallis test is robust to violations of homogeneity of variance with unequal  $n$ . Either  $F^*$  or  $W$ , or robust forms of the Kruskal–Wallis test, are preferable in this situation. Third, for some distributions, the  $F$  test is more powerful than the Kruskal–Wallis test, whereas for other distributions, the reverse is true. Thus neither approach is always better than the other. Fourth, level of measurement continues to be controversial as a factor that might or might not influence the choice between parametric and nonparametric approaches.

### EXTENSION 3E3: TWO OTHER APPROACHES: RANK TRANSFORMATIONS AND M ESTIMATORS

As if the choice between parametric and nonparametric were not already complicated, there are yet other possible techniques for data analysis, even in the relatively simple one-way, between-subjects design. As we stated at the beginning of Extension 3E2, statisticians are constantly inventing new methods of data analysis. In this section, we take a brief glimpse at two methods that are still in the experimental stages of development. Because the advantages and disadvantages

of these methods are largely unexplored, we would not recommend as of this writing that you use these approaches as your sole data-analysis technique without first seeking expert advice. Nevertheless, we believe that it is important to expose you to these methods because they represent the types of innovations currently being studied. As such, they may become preferred methods of data analysis during the careers of those of you who are reading this book as students.

The first innovation, called a *rank transformation approach*, has been described as a bridge between parametric and nonparametric statistics by its primary developers, Conover and Iman (1981). The rank transformation approach consists of simply replacing the observed data with their ranks and then applying the usual parametric test. Conover and Iman (1981) discuss how this approach can be applied to such diverse problems as multiple regression, discriminant analysis, and cluster analysis. In the case of the one-way, between-subjects design, the parametric  $F$  computed on ranks (denoted  $F_R$ ) is closely related to the Kruskal–Wallis test. Conover and Iman show that  $F_R$  is related to the Kruskal–Wallis  $H$  with the formula

$$F_R = [H / (a - 1)] / [(N - 1 - H) / (N - a)].$$

The rank transformation test compares  $F_R$  to a critical  $F$  value, whereas the Kruskal–Wallis test compares  $H$  to a critical  $\chi^2$  value. Both methods are large-sample approximations to the true critical value. Iman and Davenport (1976) found the  $F$  approximation to be superior to the  $\chi^2$  approximation in the majority of cases they investigated (see Delaney & Vargha, 2002, for discussion of a situation where using rank transformations did not work well).

A second innovation involves a method of parameter estimation other than least squares. Least squares forms the basis for comparing models in all parametric techniques we discuss in this book. In one form or another, we generally end up finding a parameter estimate  $\hat{\mu}$  to minimize an expression of the form  $\Sigma(Y - \hat{\mu})^2$ . Such an approach proves to be optimal when distributions are normal with equal variances. However, as we have seen, optimality is lost when these conditions do not hold. In particular, least squares tends to perform poorly in the presence of outliers (i.e., extreme scores) because the squaring function is very sensitive to extreme scores. For example, consider the following five scores: 5, 10, 15, 20, 75. If we regard these five observations as a random sample, we could use least squares to estimate the population mean. It is easily verified that  $\hat{\mu} = 25$  minimizes  $\Sigma(Y - \hat{\mu})^2$  for these data. As we know, the sample mean, which here equals 25, is the least-squares estimate. However, only one of the five scores is this large. The sample mean has been greatly influenced by the single extreme score of 75. If we are willing to assume that the population distribution is symmetric, we could also use the sample median as an unbiased estimator of the population mean.<sup>10</sup> It is obvious that the median of our sample is 15, but how does this relate to least squares? It can be shown that the median is the estimate that minimizes the sum of the absolute value of errors:  $\Sigma|Y - \hat{\mu}|$ . Thus the sample mean minimizes the sum of squared errors, whereas the sample median minimizes the sum of absolute errors. The median is less sensitive than the mean to outliers—for some distributions, this is an advantage, but for others, it is a disadvantage. In particular, for heavy-tailed distributions, the median’s insensitivity to outliers makes it superior to the mean. However, in a normal distribution, the median is a much less efficient estimator than is the mean. The fact that neither the median nor the mean is uniformly best has prompted the search for alternative estimators.

Statisticians developed a class of estimators called M estimators that in many respects represent a compromise between the mean and the median. For example, one member of this class (the Huber M estimator) is described as acting “like the mean for centrally located observations and like the median for observations far removed from the bulk of the data” (Wu, 1985, p. 339). As a consequence, these robust estimators represent another bridge between parametric

and nonparametric approaches. These robust estimators are obtained once again by minimizing a term involving the sum of errors. However, M estimators constitute an entire class of estimators defined by minimizing the sum of some general function of the errors. The form of the function determines the specific estimator in the general class. For example, if the function is the square of the error, the specific estimation technique is least squares. Thus least-squares estimators are members of the broad class of M estimators. The median is also a member of the class because it involves minimizing the sum of a function of the errors, with the particular function being the absolute value function.

Although quite a few robust estimators have been developed, we describe only an estimator developed by Huber because of its relative simplicity.<sup>11</sup> Huber's estimator requires that a robust estimator of scale (i.e., dispersion or variability) has been calculated prior to determining the robust estimate of location (i.e., population mean). Note that the scale estimate need not actually be based on a robust estimator; however, using a robust estimator of scale is sensible, if one believes that a robust estimator of location is needed in a particular situation. Although a number of robust estimators of scale are available, we present only one: the median absolute deviation (MAD) from the median. MAD is defined as  $MAD = \text{median} \{|Y_i - Mdn|\}$ , where  $Mdn$  is the sample median. Although at first reading, the definition of MAD may resemble double-talk, its calculation is actually very straightforward. For example, consider again our hypothetical example of five scores: 5, 10, 15, 20, and 75. As we have seen, the median of these scores is 15, so we can write  $Mdn = 15$ . Then the absolute deviations are given by  $|5 - 15| = 10$ ,  $|10 - 15| = 5$ ,  $|15 - 15| = 0$ ,  $|20 - 15| = 5$ , and  $|75 - 15| = 60$ . MAD is defined to be the median of these five absolute deviations, which is 5 in our example.<sup>12</sup> MAD can be thought of as a robust type of standard deviation. However, the expected value of MAD is considerably less than  $\sigma$  for a normal distribution. For this reason, MAD is often divided by .6745, which puts it on the same scale as  $\sigma$  for a normal distribution. We let  $S$  denote this robust estimate of scale, so we have  $S = MAD/.6745$ .

With this background, we can now consider Huber's M estimator of location. To simplify our notation, we define  $u_i$  to be  $(Y_i - \hat{\mu})/S$ , where  $S$  is the robust estimate of scale (hence we already know its value) and  $\hat{\mu}$  is the robust estimate of location whose value we are seeking. Then Huber's M estimator minimizes the sum of a function of the errors  $\sum_{i=1}^n f(u_i)$ , where the function  $f$  is defined as follows:

$$f(u_i) = \begin{cases} \frac{1}{2}u_i^2 & \text{if } |u_i| \leq 1 \\ |u_i| - \frac{1}{2} & \text{if } |u_i| > 1 \end{cases}.$$

Notice that function  $f$  involves minimizing sums of squared errors for errors that are close to the center of the distribution but involves minimizing the sum of absolute errors for errors that are far from the center. Thus, as our earlier quote from Wu indicated, Huber's estimate really does behave like the mean for observations near the center of the distribution but like the median for those farther away. At this point, you may be wondering how the  $\hat{\mu}$  that minimizes the sum of Huber's function is determined. It turns out that the value must be determined through an iterative procedure. As a first step, a starting value for  $\hat{\mu}$  is chosen; a simple choice for the starting value would be the sample median. We might denote this value  $\hat{\mu}_0$ —the zero subscript indicating that this value is the optimal value after zero iterations. Then a new estimate is computed that minimizes the function  $\sum_{i=1}^n f(u_i)$ , where  $u_i = (Y_i - \hat{\mu}_0)/S$ . This yields a new estimate  $\hat{\mu}_1$ , where the subscript 1 indicates that one iteration has been completed. The process continues until it converges, meaning that further iterations would make no practical difference in the value.<sup>13</sup>

Not only does M estimation produce robust estimates, but it also provides a methodology for hypothesis testing. Schrader and Hettmansperger (1980) show how full and restricted models based on M estimates can be compared to arrive at an  $F$  test using the same basic logic that underlies the  $F$  test with least squares. Li (1985) and Wu (1985) describe how M estimation can be applied to robust tests in regression analysis.

In summary, we have seen two possible bridges between the parametric and nonparametric approaches. It remains to be seen whether either of these bridges will eventually span the gap that has historically existed between proponents of parametrics and proponents of nonparametrics.

### EXTENSION 3E4: WHY DOES THE USUAL $F$ TEST FALTER WITH UNEQUAL $N$ S WHEN POPULATION VARIANCES ARE UNEQUAL?

Why is the  $F$  test conservative when large sample sizes are paired with large variances, yet liberal when large sample sizes are paired with small variances? The answer can be seen by comparing the expected values of  $MS_W$  and  $MS_B$  when the null hypothesis is true, but variances are possibly unequal. In this situation, the expected values of both  $MS_B$  and  $MS_W$  are weighted averages of the  $a$  population variances. However, sample sizes play different roles in the two weighting schemes.

Specifically, it can be shown that if the null hypothesis is true,  $MS_B$  has an expected value given by

$$\mathcal{E}(MS_B) = \frac{\sum_{j=1}^a w_j \sigma_j^2}{\sum_{j=1}^a w_j}, \quad (3E4.1)$$

where  $w_j = N - n_j$ . Thus the weight a population variance receives in  $MS_B$  is inversely related to its sample size. Although this may seem counterintuitive, it helps to realize that  $MS_B$  is based on  $\bar{Y}_j - \bar{Y}$ , and larger groups contribute proportionally more to  $\bar{Y}$ .

Similarly, it can be shown that  $MS_W$  has an expected value equal to

$$\mathcal{E}(MS_W) = \frac{\sum_{j=1}^a w_j^* \sigma_j^2}{\sum_{j=1}^a w_j^*}, \quad (3E4.2)$$

where  $w_j^* = n_j - 1$ . Thus the weight a population variance receives in  $MS_W$  is directly related to its sample size.

What are the implications of Equations 3E4.1 and 3E4.2? Let's consider some special cases.

#### Case I. Homogeneity of Variance

If all  $\sigma_j^2$  are equal to each other, Equations 3E4.1 and 3E4.2 simplify to  $\mathcal{E}(MS_B) = \sigma^2$  and  $\mathcal{E}(MS_W) = \sigma^2$ , because the weights are irrelevant when all the numbers to be averaged are identical. In this case, the  $F$  ratio of  $MS_B$  to  $MS_W$  works appropriately, regardless of whether the sample sizes are equal or unequal.

### Case II. Unequal Variances but Equal $n$

If all  $n_j$  are equal to each other, Equations 3E4.1 and 3E4.2 simplify to  $\mathcal{E}(MS_B) = \Sigma_{j=1}^a \sigma_j^2 / a$  and  $\mathcal{E}(MS_W) = \Sigma_{j=1}^a \sigma_j^2 / a$ . Because the weights are equal to one another, in both cases, the weighted averages become identical to simple unweighted averages. Thus  $MS_B$  and  $MS_W$  are equal to one another in the long run. Although the ANOVA assumption has been violated, the  $F$  test is typically only slightly affected here.

### Case III. Unequal Variances: Large Samples Paired with Small Variances

In this situation, we can see from Equation 3E4.1 that  $\mathcal{E}(MS_B)$  receives more weight from the smaller samples, which have larger variances. Thus the weighted average used to calculate  $\mathcal{E}(MS_B)$  is larger than the unweighted average of the  $\sigma_j^2$  terms. However,  $\mathcal{E}(MS_W)$  receives more weight from the larger samples, which have smaller variances. Thus the weighted average used to calculate  $\mathcal{E}(MS_W)$  is smaller than the unweighted average of the  $\sigma_j^2$  terms. As a consequence,  $\mathcal{E}(MS_B) > \mathcal{E}(MS_W)$ , even when the null hypothesis is true.  $F$  values tend to be too large, resulting in too many rejections of the null hypothesis when it is true. Thus the Type I error rate is too high.

### Case IV. Unequal Variances: Large Samples Paired with Large Variances

This situation is just the opposite of Case III. Now  $\mathcal{E}(MS_B)$  gives more weight to the groups with small variances because they are smaller in size. In contrast,  $\mathcal{E}(MS_W)$  gives more weight to the groups with large variances because they are larger in size. As a result,  $\mathcal{E}(MS_B) < \mathcal{E}(MS_W)$  when the null hypothesis is true. The  $F$  test is conservative and rejects the null hypothesis too infrequently. Thus power suffers.

## EXERCISES

- \*1. True or False: Although the parametric modification  $F^*$  is more robust than the usual  $F$  test to violations of homogeneity of variance in between-subjects designs, the  $F^*$  test is always at least slightly less powerful than the  $F$  test.
2. True or False: The parametric test based on Welch's  $W$  statistic can be either more or less powerful than the usual  $F$  test in equal- $n$  designs.
3. True or False: When sample sizes are unequal and heterogeneity of variance is suspected in one-way between-subjects designs, either  $F^*$  or  $W$  should seriously be considered as a replacement for the usual  $F$  test.
4. True or False: If one is willing to assume distributions have identical shapes, the Kruskal–Wallis test can be regarded as testing a “shift” hypothesis in location without requiring an assumption that scores are distributed normally.
5. True or False: The nonparametric Kruskal–Wallis test and the parametric  $F$  test always test the same hypothesis, but they require different distributional assumptions.
6. True or False: Although the  $F$  test is more powerful than the Kruskal–Wallis test when the normality and homogeneity of variance assumptions are met, the Kruskal–Wallis test can be more powerful than the  $F$  test when these assumptions are not met.

- \*7. True or False: When sample sizes are unequal and heterogeneity of variance is suspected in one-way between-subjects designs, the nonparametric Kruskal–Wallis test should be considered seriously as a replacement for the usual  $F$  test.
- \*8. How do the values of  $F$ ,  $F^*$ , and  $W$  compare to each other when samples are of different sizes and variances are considerably different from one another? Consider the following summary statistics:

Group 1	Group 2	Group 3
$n_1 = 20$	$n_2 = 20$	$n_3 = 50$
$\bar{Y}_1 = 10$	$\bar{Y}_2 = 12$	$\bar{Y}_3 = 14$
$s_1^2 = 10$	$s_2^2 = 10$	$s_3^2 = 50$

- Calculate an observed  $F$  value for these data.
  - Calculate the  $F^*$  value for these data (however, you need not compute the denominator degrees of freedom).
  - Calculate the  $W$  value for these data.
  - Are your answers to parts a–c consistent with the assertion made in Table 3E2.1 that when large samples are paired with large variances the  $F$  is conservative, whereas  $F^*$  and  $W$  are more robust?
9. Suppose that, as in Exercise 8, samples are of different sizes, and variances are considerably different from each other. Now, however, the large variance is paired with a small sample size:

Group 1	Group 2	Group 3
$n_1 = 20$	$n_2 = 20$	$n_3 = 50$
$\bar{Y}_1 = 10$	$\bar{Y}_2 = 12$	$\bar{Y}_3 = 14$
$s_1^2 = 50$	$s_2^2 = 10$	$s_3^2 = 10$

- Calculate an observed  $F$  value for these data.
  - Calculate the  $F^*$  value for these data (however, you need not compute the denominator degrees of freedom).
  - Calculate the  $W$  value for these data (however, you need not compute the denominator degrees of freedom).
  - Are your answers to parts a–c consistent with the assertion made in Table 3E2.1 that when large samples are paired with small variances, the  $F$  is liberal, whereas  $F^*$  and  $W$  are more robust?
  - Are the  $F$ ,  $F^*$ , and  $W$  values of this exercise higher or lower than the corresponding  $F$ ,  $F^*$ , and  $W$  values of Exercise 8? Is the direction of change consistent with Table 3E2.1?
- \* 10. Assume the following data are from a one-way between-subjects design (these data are also used in Exercise 16 at the end of Chapter 5):

Group 1	Group 2	Group 3
48	59	68
54	46	62
47	49	53
54	63	59
62	38	67
57	58	71

Perform a nonparametric test of the difference among these three groups.

## NOTES

1. Wilcox et al.'s (1986) results suggest that Type I error rates are more likely to be excessive as the number of groups increases. For example, with equal  $n$ s as small as 11, the Type I error rate of the  $t$  test remains close to .05, even when the population standard deviations have a 4:1 ratio. However, the Type I error rate for a four-group ANOVA with equal  $n$ s of 11 was .109 when the population standard deviation of one group was four times larger than the standard deviation of the other groups. Even for equal  $n$ s of 50, the Type I error rate for ANOVA was .088 in this situation. Thus, for more than two groups, wide disparities in population standard deviations can make the usual ANOVA excessively liberal, even with equal  $n$ .
2. SPSS provides the value of the test statistic, degrees of freedom, and  $p$  value for both tests.
3. Strictly speaking,  $MS_B$  and  $MS_W$  are both unbiased estimators of the same population variance if homogeneity holds and the null hypothesis is true. The further assumptions of normality and independence guarantee that the ratio of  $MS_B$  and  $MS_W$  follows an  $F$  distribution.
4. Monte Carlo studies by necessity investigate power differences only under a limited set of conditions. Nevertheless, the value of .03 would seem to be a reasonable figure for most practical situations. The single exception is likely to be where  $n$  is very small, in which case  $F$  might enjoy a larger advantage over  $F^*$ .
5. Cleveland (1985, pp. 135–143) presents two graphical techniques that are especially appropriate for judging whether the data conform to a shift hypothesis when comparing the distributions of two groups. The percentile comparison graph is obtained by plotting the percentiles of one distribution against the corresponding percentiles of the other distribution. If a shift of location describes the difference between the groups, the resultant plot should resemble a straight line. The Tukey sum-difference graph plots sums of corresponding percentiles against differences of corresponding percentiles and should resemble a flat straight line when the shift hypothesis holds. Cleveland argues that comparing means may be misleading when the percentile comparison graph is curved (or the Tukey sum-difference graph is not flat). Indeed, in such a situation, any single value (e.g., mean or median) may hide important characteristics of the difference between the two distributions. Darlington's (1973) ordinal dominance curve methodology provides an interesting alternative in this situation.
6. When population distributions have different shapes, alternative methods have been developed for testing differences between population medians. For further information, see Fligner and Rust (1982) and Wilcox and Charlin (1986).
7. Although the general consensus among statisticians is that the  $F$  test is robust to violations of non-normality, there are some dissenters to this view. For an example, the interested reader should consult Bradley (1978), who provides a very readable set of arguments for why he believes that the robustness of parametric tests has been oversold.
8. As stated earlier, Tomarken and Serlin only sampled from normal populations. Clinch and Keselman (1982) found  $F^*$  to be somewhat more robust than  $W$  when sampling from nonnormal populations.
9. Relative efficiency as sample sizes approach infinity is referred to as *asymptotic relative efficiency*, which is often abbreviated ARE. Although ARE is a useful concept, the relative efficiency of two tests in small samples may differ considerably from the ARE. In particular, one limitation of the Kruskal–Wallis test is that it is typically impossible to establish a critical value that will set  $\alpha = .05$ , even when all assumptions have been met. Especially with small sample sizes,  $\alpha$  may have to be set considerably below .05, which inevitably results in a loss of power. In such a situation, the relative efficiency of the nonparametric test suffers relative to the parametric test.
10. The sample median is always a median-unbiased estimator of the population median for random sampling. When the population distribution is symmetric, its mean and median are identical. Although the sample mean and sample median are generally different, both are unbiased estimators of the population mean of a symmetric distribution.
11. To simplify even further, we show Huber's estimator with a fixed tuning constant set equal to 1.0. See Hoaglin, Mosteller, and Tukey (1983); Huber (1981); and Wu (1985) for more details.

12. The median can be defined in more than one way when some scores are tied. We have chosen the simplest definition here, which simply ignores the presence of ties and defines the median to equal the value of the middle observation.
13. Hoaglin et al. (1983) show that M estimators can be thought of as weighted averages of the observations. Specific members of the class differ in terms of how they weight the observations. For example, the mean weights each observation equally, whereas Huber's M estimator weights observations near the center of the data more heavily than observations at the extremes.

## REFERENCES: CHAPTER 3 EXTENSIONS

- Baker, B. O., Hardyck, C. D., & Petrinovich, L. F. (1966). Weak measurements vs. strong statistics: An empirical critique of S. S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement*, 26, 291–309.
- Blair, R. C. (1981). A reaction to "Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance". *Review of Educational Research*, 51, 499–507.
- Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice Hall.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144–152.
- Brown, M. B., & Forsythe, A. B. (1974). The ANOVA and multiple comparisons for data with heterogeneous variances. *Biometrics*, 30, 719–724.
- Cleveland, W. S. (1985). *The elements of graphing data*. Belmont, CA: Wadsworth.
- Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Clinch, J. J., & Keselman, H. J. (1982). Parametric alternatives to the analysis of variance. *Journal of Educational Statistics*, 7, 207–214.
- Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35, 124–129.
- Darlington, R. B. (1973). Comparing two groups by simple graphs. *Psychological Bulletin*, 79, 110–116.
- Davison, M. L., & Sharma, A. R. (1988). Parametric statistics and levels of measurement. *Psychological Bulletin*, 104, 137–144.
- Delaney, H. D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinarily scaled variables and small to moderate sized samples. *Psychological Methods*, 7, 485–503.
- Fligner, M. A., & Rust, S. W. (1982). A modification of Mood's median test for the generalized Behrens—Fisher problem. *Biometrika*, 69, 221–226.
- Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, 87, 564–567.
- Gardner, P. L. (1975). Scales and statistics. *Review of Educational Research*, 45, 43–57.
- Gibbons, J. D. (1971). *Nonparametric statistical inference*. New York: McGraw-Hill.
- Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the analysis of variance and covariance. *Review of Educational Research*, 42, 237–288.
- Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). Introduction to more refined estimators. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 283–296). New York: Wiley.
- Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*. New York: Wiley.
- Huber, P. J. (1981). *Robust statistics*. New York: Wiley.
- Iman, R. L., & Davenport, J. M. (1976). New approximations to the exact distribution of the Kruskal-Wallis test statistic. *Communications in Statistics, Series A*, 5, 1335–1348.
- Iman, R. L., Quade, D., & Alexander, D. (1975). Exact probability levels for the Kruskal-Wallis test. In H. L. Harter & D. B. Owen (Eds.), *Selected tables in mathematical statistics* (pp. 329–384). Providence, RI: American Mathematical Society.
- Kendall, M. G., & Stuart, A. (1973). *The advanced theory of statistics: Vol. 2. Inference and relationship* (3rd ed.). London: Griffin.
- Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in the analysis of variance. *Psychological Bulletin*, 99, 422–431.
- Keselman, H. J., Rogan, J. C., & Feir-Walsh, B. J. (1977). An evaluation of some nonparametric and parametric tests for location equality. *British Journal of Mathematical and Statistical Psychology*, 30, 213–221.

- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks. *Journal of the American Statistical Association*, *47*, 583–621.
- Labovitz, S. (1967). Some observations on measurement and statistics. *Social Forces*, *46*, 151–160.
- Li, G. (1985). Robust regression. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Exploring data tables, trends, and shapes* (pp. 281–343). New York: Wiley.
- Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*, *8*, 750–751.
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*, 50–60.
- Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, CA: Brooks/Cole.
- Marcus-Roberts, H. M., & Roberts, F. S. (1987). Meaningless statistics. *Journal of Educational Statistics*, *12*, 383–394.
- Maxwell, S. E., & Delaney, H. D. (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin*, *97*, 85–93.
- Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? *American Psychologist*, *70*, 487–498.
- Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.
- Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, *100*, 398–407.
- Noether, G. E. (1976). *Introduction to statistics: A nonparametric approach*. Boston: Houghton Mifflin.
- Oshima, T. C., & Algina, J. (1992). Type I error rates for the James's second order test and Wilcox's  $H_m$  test under heteroscedasticity and nonnormality. *British Journal of Mathematical and Statistical Psychology*, *45*, 255–263.
- Rogan, J. C., & Keselman, H. J. (1977). Is the ANOVA  $F$ -test robust to variance heterogeneity when sample sizes are equal? An investigation via a coefficient of variation. *American Educational Research Journal*, *14*, 493–498.
- Schrader, R. M., & Hettmansperger, T. P. (1980). Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika*, *67*, 93–101.
- Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*, 73–83.
- Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.
- Spencer, B. D. (1983). Test scores as social statistics: Comparing distributions. *Journal of Educational Statistics*, *8*, 249–269.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 667–680.
- Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.
- Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*, 90–99.
- Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin*, *96*, 394–401.
- Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics*, *23*, 195–217.
- Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330–336.
- Wilcox, R. R. (1985). On comparing treatment effects to a standard when the variances are unknown and unequal. *Journal of Educational Statistics*, *10*, 45–54.
- Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.
- Wilcox, R. R., & Charlin, V. L. (1986). Comparing medians: A Monte Carlo study. *Journal of Educational Statistics*, *11*, 263–274.
- Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA  $F$ ,  $W$ , and  $F^*$  statistics. *Communications in Statistics-Simulation and Computation*, *15*, 933–943.
- Wu, L. L. (1985). Robust  $m$ -estimation of location and regression. In N. B. Tuma (Ed.), *Sociological methodology* (pp. 316–388). San Francisco: Jossey-Bass.
- Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, *57*, 173–181.

