Nonparametric Statistics

The term "parametric" is intended to refer to statistical tests that make assumptions about particular population parameters (e.g., equal variances in two groups in the population) or use particular distributions for making statistical decisions (e.g., use of the t distribution). The term "nonparametric" is intended to refer to statistical tests that do not make some or any of the standard assumptions of parametric tests. The distinction is not so clear cut. Sheskin (2011), author of a large tome on nonparametric tests, states "In truth, nonparametric tests are really not assumption free, and, in view of this, some sources Marascuilo and McSweeney (1977) suggest that it might be more appropriate to employ the term 'assumption freer' rather than nonparametric in relation to such tests" (p. 109).

Researchers sometimes turn to nonparametric statistical tests when they suspect that parametric statistical tests, such as the t test or ANOVA, have unequal (heterogeneous) variances or the dependent variable is nonnormal. Keep in mind that these assumptions are about the population and that, although the sample may suggest variance differences or nonnormality, it is not certain that the population will have these characteristics. Keep in mind also that the central limit theorem shows that the sampling distribution will be approximately normal even when the population is severely nonnormal. As shown in Table 6.3 of your main text (Myers, Well, & Lorch, 2010, p. 137), the standard unadjusted t test fairs pretty well in the face of unequal variances except when sample size is small, variances are very unequal (e.g., a 4-to-1 ratio), and sample sizes are unequal. For non-experimental, field research, and applied work in the social sciences, these circumstances may occasionally arise, but they are not extremely common in my experience. Remember also that one parametric test adjustment, the Welch's adjustment ("equal variances not assumed" in SPSS) for t tests or ANOVA, can mitigate some of the problems arising with unequal variances (Algina, Oshima, & Lin, 1994).[1]

Another justification often cited for using nonparametric tests is that ordinal variables, such as those derived from Likert scales, are not truly continuous. I have addressed this issue in other handouts (see "Ordinal Analyses" and "Levels of Measurement and Choosing the Correct Statistical Test"). There are a number of simulation studies that suggests that if there are five or more ordered categories, there will be relatively little harm in treating these ordinal variables as continuous under many conditions (e.g., Johnson & Creech, 1983; Muthén & Kaplan, 1985; Zumbo & Zimmerman, 1993; Taylor, West, & Aiken, 2006).

Considering all of the above, there are still some circumstances in which nonparametric test are preferred over their much more commonly used parametric counterparts. Before we discuss specifics, let's discuss how we would evaluate whether one test "performs better" than another test.

Relative Efficiency of a Test

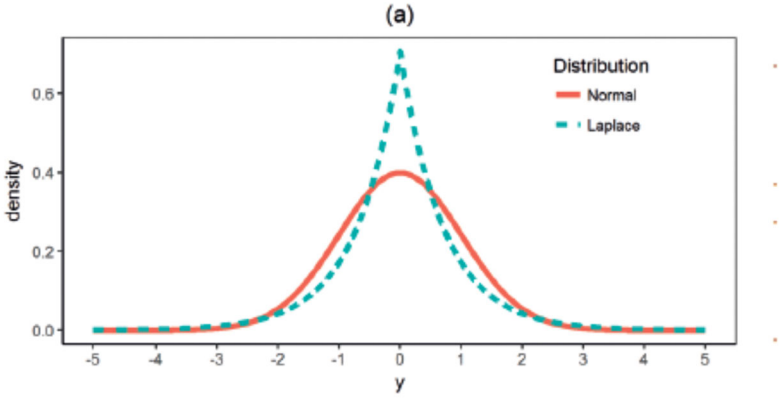
Both Type I and Type II errors are potential concerns when assumptions of parametric tests are not met. Statisticians and researchers generally seek to find a test that provides the lowest of both types of errors as well as a test that performs the best under the most common circumstances. When a statistical value (e.g., the mean) has a smaller standard error (or, more commonly, the variance of the sampling distribution) compared with another test (e.g., median), it is said to have a higher *relative efficiency*. In general, any statistic is often referred to by the lowercase Greek theta, $\hat{\theta}$ (the caret symbol signifies a sample estimate), and the relative efficiency is computed as a ratio of squared deviations of the statistic from the population value, θ.

$$RE_{1to2} = \frac{E\left(\hat{\theta}_2 - \theta\right)^2}{E\left(\hat{\theta}_1 - \theta\right)^2}$$

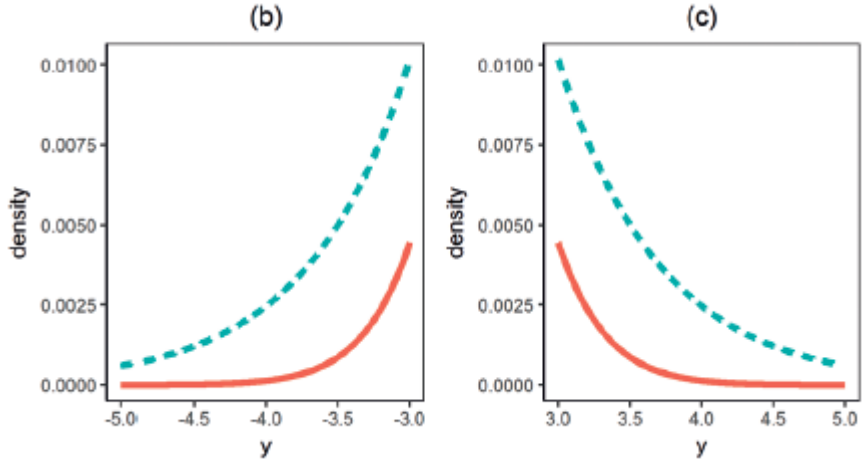
[1] The James test (or James's test) is also a possible alternative. Algina and colleagues found the Welch test to be less sensitive to unequal variances but may have more problematic Type I errors when skewness is present. Other studies, such as that by Krishnamoorthy, Lu, and Mathew (2007) suggest that the James test may be preferred under the conditions they studied (unequal variances, unequal n's, and skewness) but not as optimal as bootstrapping.

Table 6.3 in Nussbaum (2014) provides an excellent summary of the circumstances in which nonparametric tests outperform parametric tests. The relative efficiency (ARE for asymptotic relative efficiency) is shown in the third column, where values greater than 1.0 indicate that the nonparametric test has greater efficiency than the parametric test. Nearly all of the conditions in which the ARE is greater than 1.0 occur when the distribution is extremely kurtotic or skewed—a double exponential distribution (also known as the Laplace distribution).[2] The double exponential form involves raising a variable to some power and then raising it again to a power. Double exponential distributions may be symmetric or asymmetric, each having a variety of forms depending on the choices of the location (center) and spread parameters. The symmetric versions are highly kurtotic and the asymmetric versions are highly skewed. In both symmetric and asymmetric forms, extreme values are much more likely, which causes conventional parametric tests to be less powerful than nonparametric tests. A couple of illustrations of the symmetric version are shown below (both from Geraci & Borja, 2018, p. 11).

Double exponential (Laplace) distribution normal (solid orange) vs. Laplace (dotted blue):



Close-up of the extreme right and left tails of the normal (solid orange) vs. Laplace (dotted blue)



Brief Summary of a Few Common Nonparametric Tests

There are a wide variety of nonparametric tests that have been developed (see Sheskin, 2011 for a comprehensive overview), and I will discuss just a few of the most common.

Comparing two independent groups. *Mann-Whitney U* and *Wilcoxon rank sum* tests are equivalent tests (SPSS prints both together and one significance test for the two statistical values, and sometimes they are referred to jointly as the *Wilcoxon-Mann-Whitney* test) that are based on the rank ordering of the cases and then a comparison much like the t test on those ranks. This test can be more powerful than the t test, but it is problematic if the variances of the two groups are unequal.

[2] After the 18th and 19th century mathematician Pierre-Simon Laplace, 1749-1827.

Comparing two or more independent groups. The *Kruskal-Wallis* or the *Kruskal-Wallis H* test is an alternative to between-subjects one-way ANOVA. The scores are ranked and ties are resolved by inserting the median of a set of similar values. Another alternative is the *rank-transform F test*. These tests can outperform standard ANOVA if the variances and distributions are similar across the groups.

Comparing two related scores. The most common within-subjects nonparametric test that corresponds with the correlated-scores or paired t test is called the *sign* test. The sign test assigns a plus or minus (two possible values) according to whether the score for a case (or pair) increases or decreases. The test then determines if there are more pluses or minuses. The sign test is equivalent to a test of whether or not the difference has a median equal to zero. If the variable is binary, the test is equivalent to the McNemar's test. The *Wilcoxon signed rank test*, in which the absolute values of the differences are ranked, is a more powerful alternative to the sign test. The paired t test has greater relative efficiency when the data are skewed or there are many difference scores equal to 0.

Comparing more than two related scores. The most commonly mentioned nonparametric alternative to within-subjects ANOVA is *Friedman's* test, which is distributed approximately as a chi-square. Similar to the Kruskal-Wallis test, scores are ranked and ties are assigned grouped median values. Another nonparametric alternative to the within-subjects ANOVA is the *rank-transformation F test*. The rank transformation test involves ranking of individual values (rather than per pair) and then a within-subjects ANOVA is computed. The test is just one of a group of tests that use parametric test on after values have been transformed to ranks. The Friedman and the rank-transformation tests may be better than within-subjects ANOVA when the distribution is highly skewed, but the relative efficiency of the two depends on a number of circumstances. Myers and colleagues (2010) recommend the rank-transformation approach when there are more than five levels.

Other Alternatives

There are several other alternatives to traditional nonparametric tests when assumptions are in doubt. One common approach is to transform variables using a nonlinear function (e.g., squaring, square root, logarithm, Box-Cox normalizing function) to improve the distribution of the dependent variable. This approach is more acceptable in some areas, such as economics, but is less favored in psychology and the social sciences because the interpretation becomes more obscured. Still, there are some measures, such as reaction time for which transformations are commonly employed (and consequently more understood). Notice that nonparametric tests that involve ranking the data are really employing a nonlinear transformation. This is particularly evident for tests such as the rank-transformation F test for within-subjects, because the parametric F test is used with the ranked values.

More recently, tests using trimmed means, in which, most often, the lowest and highest 5% of scores are trimmed from each group, have gained popularity (e.g., Tukey & McLaughlin, 1963; Yuen, 1974). Trimming, as you might expect, reduces the variance and, consequently, reduces the standard error. This approach can have advantages over parametric tests when data are heavily skewed or kurtotic (e.g., Doksum & Wong, 1983). The parametric t test or ANOVA is then used with the trimmed data. The disadvantages to trimmed approaches are that a proportion of the data are lost and potentially informative or valuable scores are not used.

Another approach that is gaining considerable credibility with some types of analyses (e.g., hierarchical linear modeling), is the use of standard error adjustments, called robust standard errors (also known as Huber-White, Eiker-Huber-White, or sandwich estimators) that can help with heterogeneity of variance issues and nonnormal distributions. The approach shows considerable promise in a number of simulations studies (e.g., Long & Ervin, 2000) but has been slow to be incorporated into simpler statistical tests such as t tests and ANOVA.

Perhaps the most popular approach to distribution problems is to use bootstrapping (Efron, 1979). Bootstrap estimates are derived from repeated resampling from the study data set with the same sample size, just retaken with replacement. Usually at least 500 to 1000 samples (replications) are recommended.

Standard errors and confidence limits can then be estimated as if the multiple samples formed a miniature sampling distribution (*non-parametric* bootstrapping uses the distribution of the samples and *parametric* bootstrapping restricts the distribution to a particular statistical distribution shape, such as normal, binomial, Poisson etc.). The approach can be very useful when the standard error is undefined, when sampling distribution is nonnormal, or there are outliers. Bootstrap estimates are not widely available in standard software packages for all types of tests, and it is not fully resolved whether and under what conditions bootstrapping has advantages over standard parametric tests. In a simulation study for comparing multiple means (usually compared with ANOVA) in small sample sizes (n per cell $\leq$ 15) when there were unequal variances, Krishnamoorthy and colleagues (Krishnamoorthy, Lu, & Mathew, 2007) showed that parametric (normal) bootstrap estimates outperformed Welch's test, the generalized F test, and the James test, particularly for small sample sizes, although the James test was close to comparable in many circumstances. Parras-Futros (2014) showed similar advantages with a nonparametric bootstrap approach.

Bayesian estimation is another potential method of addressing analysis challenges with small sample sizes and distributional problems. Bayesian estimation does not make standard distributional assumptions per se, but derives solutions based on prior distributions which can be based on prior research, theory, or tailored to the observed data. An advantage is that the Bayesian approach depends more on the number of sampling draws (assuming the commonly used Markov chain Monte Carlo or MCMC estimation process) than on the sample size (Kadane, 2015). A challenge for the Bayesian approach, however, is that high quality prior information about the distribution is especially needed with small samples. Accurate results with smaller sample sizes are reliant on informative (strong) and good prior values (van de Schoot et al., 2014). Caution is especially needed for small samples when using diffuse (weak, or noninformative) priors (Gelman, 2006). Without careful attention to priors and judicial implementation, Bayesian estimation with small samples can lead to worse estimates than traditional or nonparametric tests (McNeish, 2016).

References

Algina, J., Oshima, T. C., & Lin, W. Y. (1994). Type I error rates for Welch's test and James's second-order test under nonnormality and inequality of variance when there are two groups. *Journal of Educational Statistics*, *19*(3), 275-291.

Doksum, K. A., & Wong, C. W. (1983). Statistical tests based on transformed data. *Journal of the American Statistical Association, 78*(382), 411-417.

Efron, B. (1979). "Bootstrap methods: Another look at the jackknife". *The Annals of Statistics. 7,* 1–26.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis, 1*, 515–534. doi:10.1214/06-BA117A

Geraci, M., & Borja, M. C. (2018). Notebook: The laplace distribution. *Significance, 15*(5), 10-11. https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1740-9713.2018.01185.x

Johnson, D.R., & Creech, J.C. (1983) Ordinal measures in multiple indicator models: A simulation study of categorization error. American Sociological Review, 48, 398-407.

Kadane, J. B. (2015). Bayesian methods for prevention research. *Prevention Science, 16*, 1017–1025. doi:10.1007/s11121-014-0531-x

Krishnamoorthy, K., Lu, F., & Mathew, T. (2007). A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models. *Computational Statistics & Data Analysis, 51*(12), 5731-5742.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician, 54*, 217-224.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 750-773.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. British Journal of Mathematical and Statistical Psychology, 38(2), 171-189.

Nussbaum, E. M. (2014). *Categorical and nonparametric data analysis: Choosing the best statistical technique*. New York: Routledge.

Parra-Frutos, I. (2014). Controlling the Type I error rate by using the nonparametric bootstrap when comparing means. *British Journal of Mathematical and Statistical Psychology, 67*(1), 117-132.

Taylor, A. B., West, S. G., & Aiken, L. S. (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. Educational and psychological measurement, 66(2), 228-239.

Sheskin, D. J. (2011). *Handbook of parametric and nonparametric statistical procedures, Fifth Edition. Boca Raton, FL*: CRC Press.

Tukey, J. W., & McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization 1. *Sankhyā: The Indian Journal of Statistics, Series A*, 331-352.

Wilcox, R. R. (1992). Comparing one-step m-estimators of location corresponding to two independent groups. Psychometrika, 57, 141-154.

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. Child Development, 85, 842– 860. doi:10.1111/cdev.12169

Yeun, K. K. (1974). The two-sample trimmed t for unequal populations. Biometrika, 61, 165-170.

Zimmerman, D. W., & Zumbo, B. D. (1993). Rank transformations and the power of the Student t test and Welch t'test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *47*(3), 523. Zumbo, B. D., & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology/Psychologie Canadienne*, *34*(4), 390.