Appendix: An overview of robust statistics

The goal of this appendix is to explain the benefits and rationale for using robust statistics and to provide a conceptual overview of these techniques. In general, robust statistical techniques are used to provide the most accurate description and analysis of a dataset. Robust techniques are designed to minimize the effects of two factors that can give rise to misleading results with traditional statistical techniques. First, most traditional statistical measures are very sensitive to the tails of a distribution (i.e., outliers; Erceg-Hurn and Mirosevich 2008; Wilcox 1995). Second, traditional inferential statistical tests are sensitive to deviations from normality, even if those deviations are small (Erceg-Hurn and Mirosevich 2008; Wilcox 1998; Wilcox and Keselman 2003). When faced with these problems, some researchers turn to traditional non-parametric tests. However, many non-parametric tests are designed for cases when the magnitude of the difference between two data points is considered not to be meaningful. For example, when asking participants to judge the loudness of a stimulus on a scale from one to seven, the perceptual distance between a 2 and a 3 rating may not be the same as the perceptual distance between a 4 and a 5 rating. Non-parametric techniques often reduce the data to a rank-based scale. When the magnitude information is meaningful, such as the difference between signal-to-noise ratios for two conditions, using such a non-parametric method will remove meaningful information.

Outliers

When there are no outliers and the data is symmetrically distributed, various measures of the central tendency, such as the mean, trimmed mean, or median, will all give the same estimate. However, the mean can be a poor description of the central tendency of a distribution when there are outliers, largely because one data point is sufficient to change the calculated mean by an arbitrarily large amount. Although most test measures impose inherent limits on the magnitude of the outliers (e.g., accuracy is bounded by 0 and 100%), preventing infinitely large shifts of the mean, outliers can still dramatically affect means, particularly with small sample sizes. An alternative robust approach for estimating the central tendency is to use trimmed means, where the mean of the central portion of the data is calculated. For most data, this will result in a value that is near the center of the bulk of the data. Although many researchers are hesitant to use trimmed means because of the perception that this technique discards data, the very large and very small values are actually converted into rank values rather than discarded. In essence, the trimmed mean is a cross between a mean and a median. To calculate the trimmed mean, all values are rank ordered and the arithmetic mean is calculated based on the middle-ranked values. For the purposes of this study and previous studies (e.g., Aronoff et al. 2014; Aronoff et al. 2011, 2012; Aronoff and Landsberger 2013; Aronoff et al. 2015), the 20% trimmed mean was calculated meaning that the largest 20% of the data and the smallest 20% of the data were converted to rank values and the mean of the central 60% of the data was calculated. This percentage was used because it has been found to provide good control in terms of Type I error for a wide range of distributions (Wilcox 1995).

Non-normality

One of the key reasons that traditional statistical technics often provide inaccurate descriptions and conclusions regarding a dataset is the often inaccurate assumption of normality in traditional (parametric) statistical techniques. There are a number of characteristics that make up a normal distribution, many of which will not be true for nearly every data set involving human data. These include:

1) Bell-shaped and unimodal
2) Symmetrical – there are an equal number of points to the left and the right of the center of the distribution
3) The mean, median, trimmed mean, and mode are all the same value
4) The tails are infinitely long – the distribution contains infinitely large and small values
5) 68% of the data is within one standard deviation of the mean
6) 95% of the data is within two standard deviations of the mean
7) 99.7% of the data is within three standard deviations of the mean

Traditional statistical techniques transform datasets into normal distributions, usually by scaling the mean and standard deviation of the normal distribution to match that of the dataset. However, deviations from normality are typical if not ubiquitous (Micceri 1989). In these cases, traditional statistical techniques can greatly distort a dataset. This can be seen in Figure A1. The top row of Figure A1 shows three distributions. The left distribution is a normal distribution. The middle distribution is a heavy-tailed distribution. For this distribution, ten percent of the normal distribution was replaced by data from a normal distribution with the same mean but a considerably larger standard deviation. This mimics what would occur when a moderate number of outliers are present. The right distribution is an asymmetric distribution similar to what would occur when using data where negative numbers are impossible, such as with the number of correct responses. When these distributions are analyzed using traditional techniques, where the data are fit to normal distributions with the same mean and standard deviation as the original data, there is a considerable distortion of the original distributions (as shown in the second row).

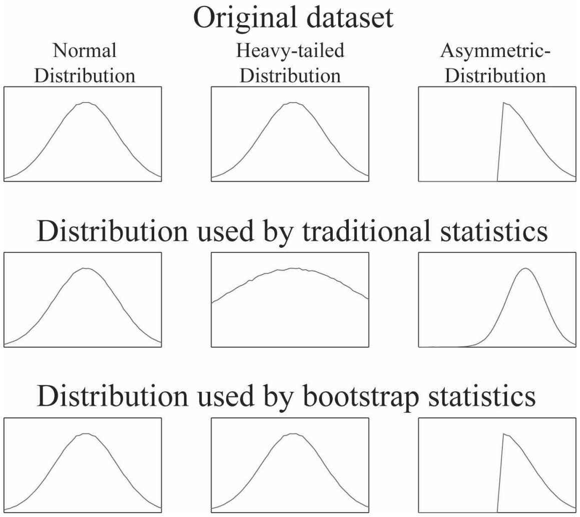

Figure A1. An example of how normal and non-normal distributions are treated with traditional analyses and bootstrap analyses.

In addition to distorting the data, even slight departures from normality can cause traditional inferential statistical methods to perform poorly, resulting in greatly reduced power (Erceg-Hurn and Mirosevich 2008; Wilcox 1998; Wilcox and Keselman 2003). Given that deviations from normality are very common in real data (Bradley 1977; Micceri 1989; Miller 1988; Wilcox 1990), even when sample sizes are large (Micceri 1989), the sensitivity of traditional statistical techniques to non-normality is not merely a theoretical concern.

One technique that can be used in conditions with normality or non-normality is the bootstrap approach (DiCiccio and Romano 1988; Efron 1982; Erceg-Hurn and Mirosevich 2008; Keselman et al. 2008; Keselman et al. 2003; Wilcox et al. 1998). Bootstrap analyses avoid the assumptions of normality by using a probability distribution based on the same characteristics as the data. This distribution, referred to as a bootstrap distribution, is generated by repeatedly sampling with replacement from the original dataset, such that the bootstrap distribution has the same number of data points as were in the original distribution. This means that the analyzed distribution matches that of the original data (see the bottom row of Figure A1).

For each bootstrap distribution, a statistical measure is calculated (such as the mean or t-statistic), resulting in a range of values for that statistical measure. Based on that range of values, a 95% confidence interval is typically calculated (i.e., a range that includes 95% of those values). This confidence interval is generally the metric used to determine significance. If the full range of the confidence interval does not include 0, then there was a significant effect. In contrast, if zero falls within the confidence interval of the mean, there was no significant effect of interleaving. A large number of bootstrap distributions are created to arrive at a stable measure (2000 in the current study). Unlike increasing the number of participants, increasing the number of bootstrap distributions does not increase the statistical power. Instead, it only increases the reliability of the estimation of the statistical metric, as shown in Figure A2.

Combining robust inferential and descriptive statistics

To deal with both non-normality and outliers, it is possible to combine robust inferential statistics with robust descriptive statistics. Bootstrap distributions are used to create a confidence interval of a statistical metric. The general bootstrap framework is agnostic as to the statistical metric used, and as such, robust descriptive statistics such as trimmed means can be used to replace traditional descriptive statistics such as means. The statistical metric chosen has no effect on the way that the bootstrap distributions are derived. The same bootstrap distributions will be used whether the descriptive statistic chosen is the mean or the trimmed mean. The only difference is whether the confidence intervals are based on the means or trimmed means of those bootstrap distributions. By combining robust descriptive statistics with robust inferential statistics, it is possible to have maximally accurate statistical measures (Wilcox et al. 1998), assuring the most accurate conclusions.

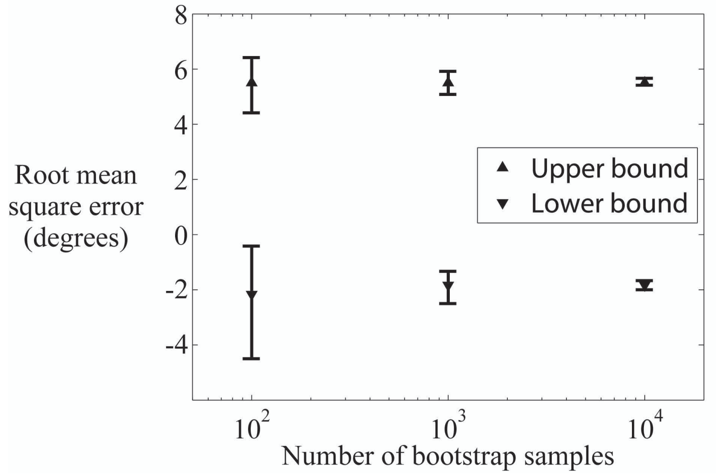

Figure A2. Median and range of upper and lower bounds of the confidence intervals for the 20% trimmed mean of the CI patients' interleaved versus non-interleaved localization scores for different numbers of bootstrap distributions. The median confidence interval does not change with increasing bootstrap distributions but the range of upper and lower bounds becomes less variable.

Robust methods for handling familywise error

When multiple comparisons are made on highly related data sets (e.g., comparing group one to group two, group one to group three, and group two to group three), the issue of familywise error arises where the probability of Type I errors increase. This arises because with the criteria for significance (α) being $p < .05$, the likelihood of obtaining a significant result by chance for one analysis is 1/20. However, if you run two analyses on highly related data sets using that criterion, the likelihood of obtaining a significant result for at least one of those analyses grows based on Equation 1.

$$\alpha = 1-(1- \alpha')^{n} \qquad \text{Equation 1}$$

where α' is corrected α (0.05 when no correction is applied) and n is the number of related tests conducted.

A standard approach to address this problem is to use a Bonferroni correction whereby α is divided by the number of tests conducted. However, Bonferroni's method controls Type II error relatively well for the case where there is only one significant result, ignoring the likelihood of multiple significant results. Rom's method (Rom 1990) takes the likelihood of multiple significant results into account. Like a Bonferroni correction, this method of mediating familywise error has good Type I error control. However, it has much better control of Type II error than a Bonferroni correction. To conduct Rom's method, all the *p* values are first rank ordered from largest to smallest. The largest *p* value is selected for the first iteration and the following steps are used:

1) The selected *p* values is compared to the adjusted α from Table A1
2) If the selected *p* value is less than the adjusted α than that value and all smaller *p* values are significant
3) If the selected *p* value is not less than the adjusted α then that value is not significant and the next largest *p* value is selected for the next iteration, starting again at step 1 (but increasing the iteration number)

This iterative process is repeated until either one of the *p* values is less than the adjusted α or there are no *p* values left to compare.

Iteration number	Adjusted α
1	0.05
2	0.025
3	0.0169
4	0.0127
5	0.0102
6	0.00851
7	0.00730
8	0.00639
9	0.00568
10	0.00511

Table A1. Corrected α for each iteration of the sequential rejection procedure (adapted from Rom 1990).

References

Aronoff, J. M., Amano-Kusumoto, A., Itoh, M., et al. (2014). The effect of interleaved filters on normal hearing listeners' perception of binaural cues. *Ear Hear, 35,* 708-710.

Aronoff, J. M., Freed, D. J., Fisher, L. M., et al. (2011). The effect of different cochlear implant microphones on acoustic hearing individuals' binaural benefits for speech perception in noise. *Ear Hear, 32*, 468-484.

Aronoff, J. M., Freed, D. J., Fisher, L. M., et al. (2012). Cochlear implant patients' localization using interaural level differences exceeds that of untrained normal hearing listeners. *J Acoust Soc Am, 131*, EL382-387.

Aronoff, J. M., Landsberger, D. M. (2013). The development of a modified spectral ripple test. *J Acoust Soc Am, 134*, EL217-222.

Aronoff, J. M., Padilla, M., Fu, Q. J., et al. (2015). Contralateral masking in bilateral cochlear implant patients: a model of medial olivocochlear function loss. *PLoS One, 10*, e0121591.

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. *The American Statistician, 31*, 147-150.

DiCiccio, T. J., Romano, J. P. (1988). A review of bootstrap confidence intervals (with discussion). *Journal of the Royal Statistical Society, B50.*

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Society for Industrial and Applied Mathematics.

Erceg-Hurn, D. M., Mirosevich, V. M. (2008). Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *Am Psychol, 63*, 591-601.

Keselman, H. J., Algina, J., Lix, L. M., et al. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychol Methods, 13*, 110-129.

Keselman, H. J., Wilcox, R. R., Lix, L. M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophysiology, 40*, 586-596.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Miller, J. (1988). A warning about median reaction time. *J Exp Psychol Hum Percept Perform, 14*, 539-543.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika, 77*, 663-666.

Wilcox, R. R. (1990). Comparing the means of two independent groups. *Biometrics Journal, 32*, 771-780.

Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research, 65,* 51-77.

Wilcox, R. R. (1998). The goals and strategies of robust methods. *British Journal of Mathematical and Statistical Psychology, 51*, 1-39.

Wilcox, R. R., Keselman, H. J. (2003). Modern robust data analysis methods: measures of central tendency. *Psychol Methods, 8*, 254-274.

Wilcox, R. R., Keselman, H. J., Kowalchuk, R. K. (1998). Can tests for treatment group equality be improved?: The bootstrap and trimmed means conjecture. *British Journal of Mathematical and Statistical Psychology, 51*, 123-134.