Journal of Applied Biomechanics, (Ahead of Print)
https://doi.org/10.1123/jab.2017-0269

Human Kinetics
ORIGINAL RESEARCH

Data Analyses When Sample Sizes Are Small: Modern Advances for Dealing With Outliers, Skewed Distributions, and Heteroscedasticity

Rand Wilcox,[1] Travis J. Peterson,[2] and Jill L. McNitt-Gray[1]
[1]University of Southern California; [2]California Lutheran University

The paper reviews advances and insights relevant to comparing groups when the sample sizes are small. There are conditions under which conventional, routinely used techniques are satisfactory. But major insights regarding outliers, skewed distributions, and unequal variances (heteroscedasticity) make it clear that under general conditions they provide poor control over the type I error probability and can have relatively poor power. In practical terms, important differences among groups can be missed and poorly characterized. Many new and improved methods have been derived that are aimed at dealing with the shortcomings of classic methods. To provide a conceptual basis for understanding the practical importance of modern methods, the paper reviews some modern insights related to why methods based on means can perform poorly. Then some strategies for dealing with nonnormal distributions and unequal variances are described. For brevity, the focus is on comparing 2 independent groups or 2 dependent groups based on the usual difference scores. The paper concludes with comments on issues to consider when choosing from among the methods reviewed in the paper.

Keywords: robust methods, distribution-free techniques, nonparametric methods, nonnormality, ANOVA

Classic, routinely taught and used statistical methods for comparing groups are based on means. From basic principles, there are 2 fundamental goals: (1) to control the probability of a type I error, meaning rejecting the null hypothesis when it is true, and (2) to have relatively high power, meaning the probability of detecting a true difference. A positive feature of standard methods for comparing means is that they control the type I error probability reasonably well when observations are sampled from populations of individuals that have identical means, variances, skewness, and so forth. That is, sampling is from identical distributions. However, recent insights indicate that under general conditions when group distributions differ these 2 fundamental goals are not achieved,[1–7] resulting in poor power and a highly misleading summary of the data. If distributions differ, even with relatively large sample sizes ($n > 200$), classic methods can yield inaccurate confidence intervals and can have relatively poor power.[6,7] In practical terms, if Student's t test fails to reject, one possibility is that there is little or no difference between the groups, or it might be because power is poor due to nonnormality (see Supplementary Material [available online]; Welch's 2-sample t test and Yuen's method with trimmed means). One strategy for dealing with nonnormal distributions is to use classic rank-based methods. They test the hypothesis that groups have identical distributions. But in terms of comparing means or medians, they are unsatisfactory under general conditions.[8,9] This remains the case when using more modern rank-based methods described by Brunner et al.[10] Another strategy is to rely on the generalized linear model. But when using means, this does not address the theoretical issues described by Huber and Ronchetti,[2] Hampel et al,[11] and Staudte and Sheather.[5] Other concerns are described by Keselman et al.[12]

Recent advances in modern statistical methods provide robust hypothesis testing by dealing with outliers and skewed distributions in ways that maintain power and control the probability of a type I error. No single method dominates in that each approach is sensitive to different features of the data and each tells something different about how groups compare. To provide a conceptual basis for understanding the practical importance of modern methods, this paper reviews some modern insights related to why methods based on means can perform poorly. Strategies for dealing with nonnormal distributions and unequal variances are also described. Modern advances for dealing with outliers, skewed distributions, limitations associated with the central limit theorem, and unequal variances (heteroscedasticity) in the context of small sample sizes (<10 participants) will be discussed. For brevity, examples will be limited to the comparison of 2 independent groups or 2 dependent groups based on the usual difference scores. The paper concludes with comments on issues to consider when choosing between the methods reviewed in the paper.

Analysis of Small Sample Sizes

Small sample sizes may be due to a very limited subject pool such as when studying elite athletes or data collection that might be difficult due to the time and expense that is required. For such situations, there is the practical issue of choosing an appropriate statistical technique. A related issue is interpreting the results in an appropriate manner. This entails having a precise understanding of what a statistical method reveals about the data and just as important, understanding what it does not reveal.

Space limitations preclude a detailed explanation regarding when and why modern methods offer a practical advantage.

Wilcox is with the Department of Psychology, University of Southern California, Los Angeles, CA, USA. Peterson is with the Department of Exercise Science, California Lutheran University, Thousand Oaks, CA, USA. McNitt-Gray is with the Department of Biomedical Engineering, University of Southern California, Los Angeles, CA, USA. McNitt-Gray is also with the Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA. McNitt-Gray (mcnitt@usc.edu) is corresponding author.

The goal here is to provide some indications regarding their relative merits. The paper begins by briefly outlining the nature of 3 major advances related to outliers, skewed distributions, and heteroscedasticity (unequal variances). Figures are included to provide context. Some unsatisfactory strategies for salvaging classic techniques are also discussed. This is followed by a section on robust methods for testing hypotheses that deal effectively with small sample sizes in a manner that addresses the issues previously described and a description of some of the more basic methods that are now available for consideration (see Supplementary Material [available online] and corresponding references). Illustrative examples from sports biomechanics research are also provided.

Outliers

The first major advance has to do with outliers and what are generally known as heavy-tailed distributions. Here it is assumed that values are flagged as outliers based on a boxplot or the so-called median absolute deviation-median rule. Methods based on the mean and variance are known to be unsatisfactory.[6] Roughly, when sampling from a heavy-tailed distribution, outliers tend to occur more frequently versus sampling from a normal distribution. This can result in poor power because outliers can inflate the standard deviation. Even a small departure from normality can result in poor power.[5–7,13] This result follows almost immediately from Tukey's[14] discussion of a mixed normal distribution.

The population variance is sensitive to slight changes in the tails of a distribution. For example, Figure 1 shows a standard normal distribution and a mixed normal distribution, which has heavy tails, meaning that the tails are thicker than the normal distribution. Despite the obvious similarity, the standard normal distribution has variance of 1, but the variance of the mixed normal is 10.9. This comparison illustrates the basic principle that the population variance is very sensitive to slight changes in the tails of a distribution.

Even when a distribution appears to be approximately normal, power can be relatively low. Figure 2 illustrates this point. The left panel shows 2 normal distributions, both of which have variances equal to 1. With sample sizes $n_1 = n_2 = 25$, power is .96 when testing at the .05 level with Student's t. By contrast, the distributions in the right panel are mixed normals. Now power is only .28 and is considered relatively low.

Put another way, outliers are more likely to occur when dealing with heavy-tailed distributions such as the mixed normal. Outliers can inflate the standard deviation to the point that power is relatively low when using means.

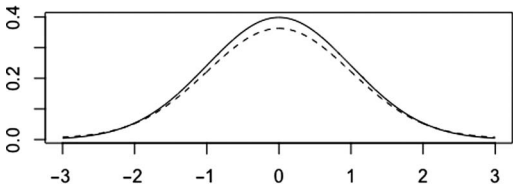

Figure 1 — The solid line is a standard normal distribution having variance 1. The dashed line is a mixed normal distribution, which is characterized as having heavy tails, meaning that the tails are thicker than a normal distribution. Despite the evident similarity, the mixed normal has variance 10.9. An implication is that even a slight departure from a normal distribution can result in relatively low power.

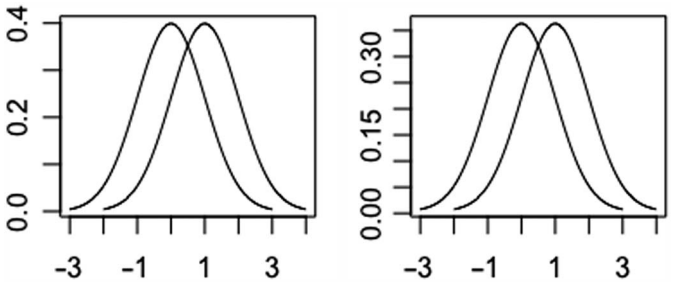

Figure 2 — In the left panel, power is .96 based on Student's t, $\alpha = .05$ and sample sizes $n_1 = n_2 = 25$. But in the right panel, the distributions are not normal, they are mixed normal, and power is only .28.

Skewed Distributions

The second advance has to do with skewed distributions, which are a much more serious concern, when using means, than is generally recognized. Suppose that for a nominal .05 type I error probability, Student's t is judged to be reasonably accurate if the actual type I probability is between .03 and .08. When observations are randomly sampled from a skewed distribution that has relatively light tails (outliers are relatively rare), approximately 130 observations can be required. When dealing with a skewed distribution where outliers are relatively common, 300 observations can be required.[7]

Skewed distributions have implications regarding the 2-sample t test. When testing at the .05 level, if the distributions being compared have the same amount of skewness, then Student's t is satisfactory in the sense that the actual type I error probability will not exceed .05. But otherwise, the actual type I error probability can be substantially larger than .05. Concerns persist regardless of how large the sample sizes might be.[15] For a more detailed summary regarding why skewed distributions are a concern, see Wilcox and Rousselet.[16]

Heteroscedasticity (Unequal Variances)

Conventional methods for comparing independent groups assume homoscedasticity (equal population variances). Heteroscedasticity (unequal population variances) can negatively impact both power and the ability to control the type I error probability. For example, when testing at the .05 level, the actual type I error probability can exceed .3 when using the analysis of variance F test.[6]

Some Unsatisfactory Strategies for Salvaging Classic Techniques

Simple transformations are sometimes suggested for salvaging methods based on means. For example, taking logarithms sometimes results in a distribution that is approximately normal. But in general, simple transformations are relatively unsatisfactory. Typically, distributions remain skewed and the deleterious impact of outliers remains.[17–19]

Another strategy is to test assumptions. For example, when comparing 2 independent groups, test the hypothesis of equal variances and if a nonsignificant result is obtained, use Student's t. However, extant publications do not support this approach.[20–24] The basic problem is that tests of the hypothesis of equal variances do not have enough power to detect situations where the equal variance assumption should be abandoned.

Dealing Effectively with Small Sample Sizes

There are robust methods for testing hypotheses that deal effectively with small sample sizes in a manner that addresses the issues previously described. Moreover, they are readily applied via the software R.[25] For the 1-sample case or when comparing dependent groups using difference scores, a few methodologies are mentioned here. The first method is based on the median, which assumes random sampling only (Supplementary Material [available online]).[6,7,9] A possible criticism is that in some situations the sample median trims too many observations. In this case, one might use a compromised amount of trimming. A 20% trimmed mean is often a good choice and has been studied extensively.[6,7] This means that after putting the observations in ascending order, the lower 20% are trimmed as well as the upper 20%. Power is nearly the same as Student's t when sampling from a normal distribution and power can be relatively high when outliers are likely to occur, as will be illustrated. It is stressed that technically sound methods for making inferences based on a trimmed mean are not obvious based on standard training. Simply trimming and using Student's t results in using an incorrect estimate of the standard error in that the observations left after trimming are dependent, which invalidates the derivation of Student's t.

In terms of type I errors, the best method uses a percentile bootstrap technique.[6,7] If the sample size is not too small, say greater than 20, the (nonbootstrap) method derived by Tukey and McLaughlin[26] is another way of dealing with trimmed means; it uses a correct estimate of the standard error. When dealing with 2 dependent groups, a third approach is to use improved versions of the sign test.[6,27]

As for comparing 2 independent groups, a method for comparing medians, via a percentile bootstrap method, performs very well in terms of controlling the type I error probability (see Supplementary Material [available online]).[28,29] Currently, it is the only known method that effectively deals with tied (duplicated) values.[7] If the likelihood of outliers is relatively low, using a 20% trimmed mean might result in better power. In addition, both theoretical results and simulation studies indicate that problems with controlling the type I error probability, when dealing with skewed distributions, diminish as the amount of trimming increases.[30]

To provide some perspective on power, consider again the 2 normal distributions in the left panel of Figure 2. As previously noted, power is .96 when comparing means at the .05 level and when both sample sizes equal to 25. Using instead a 20% trimmed mean, power is .91, and it is .78 when comparing the medians. Now consider the right panel and recall that power is .28 when comparing means. By contrast, when using a 20% trimmed mean or median, power is .75 and .70, respectively. This illustrates the general principle that when testing hypotheses based on means, power can be relatively low when outliers tend to occur.

Let P be the probability that a randomly sampled observation from the first group is less than a randomly sampled observation from the second. The classic Wilcoxon–Mann–Whitney (WMW) test is based on a direct estimate of P. But under general conditions it is not a satisfactory test of the hypothesis that $P = .5$ and it does not provide a satisfactory confidence interval for P.[8] The reason is that it uses the wrong standard error when distributions differ. In addition, under general conditions, the WMW does not test the hypothesis of equal medians as is sometimes claimed.[31] There are several methods for testing the hypothesis that $P = .5$ that continue to perform well when distributions differ in shape. Based on results in Neuhäuser et al,[32] the method derived by Cliff[8] performs particularly well when the sample sizes are small. Unlike the WMW test, Cliff's method is designed to perform well when there is heteroscedasticity.

Choosing a Method: Different Methods Provide Different Information

Several methods have been mentioned regarding how to compare 2 independent groups when the sample size is relatively small. This raises a natural question: which one is best? Because different methods are sensitive to different features of the data, the only certainty is that no single method dominates in terms of power. In addition, different methods provide different perspectives. If, for example, distributions are skewed, comparing the medians is not the same as comparing 20% trimmed means or testing the hypothesis that $P = .5$. In some situations, multiple perspectives might provide a deeper and more nuanced understanding of how groups compare.

As for comparing dependent groups, again different methods provide different perspectives. Modern variations of the sign test might seem relatively uninteresting because in some sense they ignore information that is utilized by methods based on a measure of central tendency. But situations are encountered where the sign test makes a practical difference as will be illustrated.

The first illustration deals with the comparison of vertical velocity generated by volleyball players (n = 12) when performing a block maneuver initiated with and without horizontal momentum. Five trials were performed under each condition. Here, to control the probability of one or more type I errors, an improvement on the Bonferroni method is used that was derived by Hochberg.[33] Testing the hypothesis that the difference scores have a mean of 0, using the usual paired t test, no significant differences are found. Using instead a 20% trimmed mean and the Tukey–McLaughlin method, 2 significant differences are found. Again, using a 20% trimmed mean but with the percentile bootstrap method, a significant difference is found on all 5 occasions. Using the median, a significant difference is found on 4 of the 5 occasions. The sign test returns 2 significant differences. The median absolute deviation-median rule[7] indicates that among the difference scores, the number of outliers for the 5 occasions is 1, 0, 2, 4, and 0. The very presence of outliers does not necessarily mean more power when using a trimmed mean or median, the only point is that the choice of method can make a practical difference. Russell et al[34] provide another example dealing with the sign test in a clinical context.

The second illustration stems from a study comparing the kinematics and reaction forces generated during a golf swing under 2 conditions.[35] Two types of shots were attempted on 4 different occasions resulting in 4 sets of difference scores. Again, the probability of one or more type I errors was controlled using Hochberg's method. On the third occasion, the median absolute deviation-median rule finds a kinematic variable with 6 outliers, 4 of which occur among the higher measures. The point is that with a 20% trimmed mean, only 3 of these 4 outliers are trimmed. Moreover, if the goal is to have the probability of one or more type I errors equal to .05, the hypothesis that the trimmed mean is 0 is not rejected. However, the sign test, as well as the hypothesis of a median difference equal to 0 is rejected as well. That is, despite the extreme amount of trimming used by the median, significant results can occur in contrast to results when using less trimming.

A similar result was obtained when comparing the peak force measures between the 2 shot conditions. For completeness, the paired t test also rejects in this particular case, despite the outliers. But why it

rejects is unclear for reasons previously reviewed. If a 20% trimmed mean or median rejects, it is reasonably certain that this has to do with a measure of central tendency that reflects the typical response. This is not necessarily the case when using the paired *t* test.

Concluding Remarks

Currently, we have the technology to get a deeper and more accurate understanding of our data. In particular, substantially improved statistical methods are available for dealing with small sample sizes. Moreover, free R software for applying modern methods is available.[6,7,36,37] Three of the books cited are aimed at an introductory statistics course and include both classic and modern robust methods. For SAS and SPSS users interested in using R, the book by Muenchen[38] might help. Although space limitations make it impossible to foster a deep appreciation of the many advances that have occurred, our hope is that this brief review will stimulate the use of modern techniques.

References

1. Heritier S, Cantoni E, Copt S, Victoria-Feser MP. *Robust Methods in Biostatistics*. New York, NY: Wiley; 2009.
2. Huber PJ, Ronchetti E. *Robust Statistics*. 2nd ed. New York, NY: Wiley; 2009.
3. Maronna RA, Martin DR, Yohai VJ. *Robust Statistics: Theory and Methods*. New York, NY: Wiley; 2006.
4. Rousseeuw PJ, Leroy AM. *Robust Regression and Outlier Detection*. New York, NY: Wiley; 1987.
5. Staudte RG, Sheather SJ. *Robust Estimation & Testing*. New York, NY: Wiley; 1990.
6. Wilcox RR. *Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction*. New York, NY: Chapman & Hall/CRC Press; 2017.
7. Wilcox RR. *Introduction to Robust Estimation and Hypothesis Testing*. 3rd ed. San Diego, CA: Academic Press; 2017.
8. Cliff N. *Ordinal Methods for Behavioral Data Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates; 1996.
9. Hettmansperger TP, McKean JW. *Robust Nonparametric Statistical Methods*. 2nd ed. New York, NY: Chapman & Hall; 2011.
10. Brunner E, Domhof S, Langer F. *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. New York, NY: Wiley; 2002.
11. Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. *Robust Statistics*. New York, NY: Wiley; 1986.
12. Keselman HJ, Othman A, Wilcox RR. Generalized linear model analyses for treatment group equality when data are non-normal. *J Mod Appl Stat Methods*. 2016;15:32–61.
13. Bakker M, Wicherts JM. Outlier removal, sum scores, and the inflation of the type I error rate in independent samples *t* tests: the power of alternatives and recommendations. *Psychol Methods*. 2014;19:409–427. PubMed ID: 24773354 doi:10.1037/met0000014
14. Tukey JW. A survey of sampling from contaminated normal distributions. In: Olkin I, Ghurye S, Hoeffding W, eds. *Contributions to Probability and Statistics*. Stanford, CA: Stanford University Press; 1960:448–485.
15. Cressie NAC, Whitford HJ. How to use the two sample *t*-test. *Biom J*. 1986;28:131–148. doi:10.1002/bimj.4710280202
16. Wilcox RR, Rousselet GA. A guide to robust statistical methods in neuroscience. *Curr Protoc Neurosci*. 2017;82(1):8.42.1–8.42.30. doi:10.1002/cpns.41
17. Rasmussen JL. Data transformation, type I error rate and power. *Br J Math Stat Psychol*. 2013;42:203–213. doi:10.1111/j.2044-8317.1989.tb00910.x
18. Doksum KA, Wong CW. Statistical tests based on transformed data. *J Am Stat Assoc*. 1983;78:411–417. doi:10.1080/01621459.1983.10477986
19. Wilcox RR, Keselman HJ. Modern robust data analysis methods: measures of central tendency. *Psychol Methods*. 2003;8:254–274. PubMed ID: 14596490 doi:10.1037/1082-989X.8.3.254
20. Hayes AF, Cai L. Further evaluating the conditional decision rule for comparing two independent means. *Br J Math Stat Psychol*. 2007;60:217–244. PubMed ID: 17971268 doi:10.1348/000711005X62576
21. Markowski CA, Markowski EP. Conditions for the effectiveness of a preliminary test of variance. *Am Stat*. 1990;44:322–326.
22. Moser BK, Stevens GR, Watts CL. The two-sample t-test versus Satterthwaite's approximate F test. *Comm Stat Theory Methods*. 1989;18:3963–3975. doi:10.1080/03610928908830135
23. Wilcox RR, Charlin V, Thompson KL. New Monte Carlo results on the robustness of the ANOVA F, W, and F* statistics. *Commun Stat Simul Comput*. 1986;15:933–943. doi:10.1080/03610918608812553
24. Zimmerman DW. A note on preliminary tests of equality of variances. *Br J Math Stat Psychol*. 2004;57:173–181. PubMed ID: 15171807 doi:10.1348/000711004849222
25. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2013. http://www.R-project.org. Accessed 2018.
26. Tukey JW, McLaughlin DH. Less vulnerable confidence and significance procedures for location based on a single sample: trimming/winsorization 1. *Sankhya A*. 1963;25:331–352.
27. Schilling M, Doi J. A coverage probability approach to finding an optimal binomial confidence procedure. *Am Stat*. 2014;68:133–145. doi:10.1080/00031305.2014.899274
28. Wilcox RR. Comparing medians. *Comput Stat Data Anal*. 2006;51:1934–1943. doi:10.1016/j.csda.2005.12.008
29. Yuen KK. The two sample trimmed *t* for unequal population variances. *Biometrika*. 1974;61:165–170. doi:10.1093/biomet/61.1.165
30. Wilcox RR. Some results on the Tukey–McLaughlin and Yuen methods for trimmed means when distributions are skewed. *Biom J*. 1994;36:259–273. doi:10.1002/bimj.4710360302
31. Fung KY. Small sample behaviour of some nonparametric multisample location tests in the presence of dispersion differences. *Stat Neerl*. 1980;34:189–196. doi:10.1111/j.1467-9574.1980.tb00701.x
32. Neuhäuser M, Lösch C, Jöckel KH. The Chen–Luo test in case of heteroscedasticity. *Comput Stat Data Anal*. 2007;51:5055–5060. doi:10.1016/j.csda.2006.04.025
33. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988;75:800–802. doi:10.1093/biomet/75.4.800
34. Russell IM, Raina S, Requejo PS, Wilcox RR, Mulroy S, McNitt-Gray JL. Modifications in wheelchair propulsion technique with speed. *Front Bioeng Biotechnol*. 2015;3:171. doi:10.3389/fbioe.2015.00171
35. Peterson T, Wilcox R, McNitt-Gray J. Angular impulse and balance regulation during the golf swing. *J Appl Biomech*. 2016;32:342–349. doi:10.1123/jab.2015-0131
36. Field A, Miles J, Field Z. *Discovering Statistics Using R*. Thousand Oaks, CA: Sage; 2012.
37. Wilcox RR. *Understanding and Applying Basic Statistical Methods Using R*. New York, NY: Wiley; 2017.
38. Muenchen RA. *R for SAS and SPSS Users*. 2nd ed. New York, NY: Springer; 2011.