REVIEW ARTICLE

Statistics in Nuclear Cardiology: So, What's the Difference? The t-Test: Pitfalls and Options in Hypothesis Testing for Comparing Differences in Means

David N. Williams, PhD[1], Kathryn A. Williams, MS[2] and Michael Monuteaux, ScD[3]

Received: June 18, 2018/Revised manuscript received: July 19, 2018/Accepted: July 30, 2018

Abstract

Decisions related to differences of the measures of central tendency of population parameters are an important part of clinical research. Choice of the appropriate statistical test is critical to avoiding errors when making those decisions. All statistical tests require that one or more assumptions be met. The t-test is one of the most widely used tools but is not appropriate when assumptions such as normality are not met, especially when small samples, <40, are used. Non-parametric tests, such as the Wilcoxon rank sum and others, offer effective alternatives when there are questions about meeting assumptions. When normality is in question, the Wilcoxon non-parametric tests offer substantially higher levels of power and an reliable alternative to the t-test.
Keywords: Assumptions, Difference of means, Mann-Whitney, Non-parametric, t-test, Violations, Wilcoxon
Ann Nucl Cardiol 2018; 4 (1): 83−87

Some of the most important studies in clinical research and decision-making are those related to differences in measures of central tendency of population parameters i.e. the mean or the median. Choice of the appropriate tool to adequately make those decisions depends on the hypothesis being tested and the assumptions that must be met to effectively use the tool. The t-test is one of the most widely used tools (1) but is sometimes not appropriate for the purpose. In this article we consider the t-test, the assumptions on which it is based, the effects of violations to those assumptions, and an alternative tool.

t-test assumptions and violations

All statistical tests require that one or more assumptions be met. Understanding those assumptions and whether and to what degree those assumptions are violated should be an important step in analysis but one that is often left un-done (2). The t-test is known to be "robust" against violation of assumptions but even with a robust tool, if the assumptions are not met, the risk of Type 1 or Type 2 error can increase substantially; a null hypothesis may be rejected in error or not rejected when it should be, respectively.

The t-test tool has three basic applications for testing difference of means: single sample, two sample and matched pair. They differ, but all are designed to test for difference of sample means.

The assumptions on which these tests are based are:
Single sample:

・Data must be continuous

doi: 10.17996/anc.18-00075

1) David N. Williams
Senior Biostatistician, Biostatistics and Research Design Core, Institutional Centers for Clinical and Translational Research, Department of Adolescent Medicine, Boston Children's Hospital, and Instructor of Pediatrics, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115, USA
E-mail: david.williams3@childrens.harvard.edu
2) Kathryn A. Williams
Senior Biostatistician, Biostatistics and Research Design Core, Institutional Centers for Clinical and Translational Research, Boston Children's Hospital, Boston, MA, USA
3) Michael Monuteaux
Senior Biostatistician, Biostatistics and Research Design Core, Institutional Centers for Clinical and Translational Research Department of Emergency Medicine, Associate Professor of Pediatrics, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115, USA Boston Children's Hospital

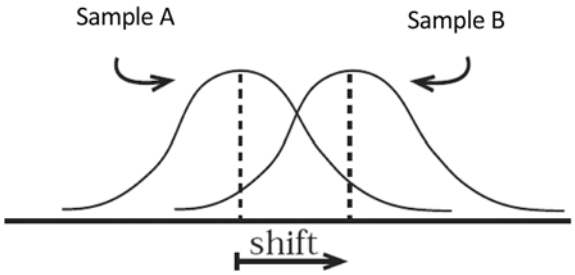

Fig. 1 Difference in the measures of central tendency of population parameters.

- Data must be normally distributed (normality)
- Simple random sample

Two samples:

- Data must be continuous
- Data must be normally distributed (normality)
- Variances of the two populations are equal (homogeneity)[1]
- The two samples are independent
- Both are simple random samples

Matched pair:

- sample sizes must be identical
- Data are continuous
- Differences for the matched pairs must be normally distributed (normality)
- The sample of pairs is a simple random sample from the population.

The assumption of normality

The least checked of these assumptions is that of normality (2). Common across all three applications (although in the matched pair application it is the differences of the matched pairs that must be normal), it is fundamentally important to the t-test. It is important because the sampling distribution of the outcome variable must be estimated (the actual population distribution is unknown); the assumption of normality allows us to make those estimates. If the assumption is correct, the distribution of the outcome variable in the population from which it was drawn is normally distributed, and a simple random sample is drawn, then any size sample will work. This is when the t-test is at its best, samples of size 20 or smaller can be used with relative immunity (although the power to detect clinically meaningful differences may be limited with small sample sizes).

But what if the assumption of normality is not being met? The t-test is often referred to as "robust (4)," meaning that moderate violations of the assumptions will not substantially affect conclusions drawn, but how does one evaluate "moderate"? It is not usually possible to determine normality of distribution for a variable in a population in a sample-based study. It is possible that research done with larger surveys or the accumulation of research findings may provide a clearer answer to the normality question. But researchers usually have to rely on examination of sample data. While statistical tests of normality are available, e. g. chi-square, Shapiro-Wilk, Anderson-Darling normality tests, they are often challenging in their applications, e.g. low power with small sample sizes, negative affect of ties (5).

A histogram presentation of the sample data (Fig. 2) will allow for a visual evaluation of normality, of skewness, the asymmetry of the variable around its mean, and kurtosis, the height and sharpness of the central peak. Comparison of basic parameters, such as mean and median, will also add insight. There are also statistical tests for skewness and kurtosis, e.g. Pearson 2 skewness coefficient or G1, if a visual assessment is not sufficient (7).

A different graphical examination of the sample data, the normal quantile-quantile (q-q) plot, can offer insight into severity of violations. This statistic is readily generated in the SAS univariate procedure. The linearity of the point pattern, such as in Fig. 3a, is an indicator that the measurements are normally distributed (8), non-linearity, Fig. 3b is an indicator of non-normality.

If the data are skewed or have extreme observations, such as in Fig. 3, then estimates of the mean may be severely affected; inferences drawn will likely be incorrect and the null hypothesis may be rejected or fail to be rejected in error (10). Another version of the q-q plot can be used to determine if two data sets come from populations with a common distribution (i.e. to test for homogeneity of variation).

What options are available if the data are severely non-normal?

1. For a two-sample t-test: A larger sample size might show normality. How much of an increase in the sample size depends on how severe the violation of normality is; but a sample size over 60 may be sufficient for moderate violations (10). Test the larger sample for normality after it is drawn.
2. Trim the data of the more extreme cases, ones that fall far from a normal distribution. Trimming must be done at specific percentile points to preserve asymptotic normality. Caution must be used when dropping extreme cases; they can represent legitimate observations (9). An outlier should only be dropped if it is due to erroneously entered or measured data or if dropping it does not change results but brings the data into compliance with assumptions (11).
3. Transform the data. This is applicable if the two samples have a similar non-normal shape. Certain transformations

[1]An un-equal variance t-test is available and is a suggested alternative to the standard t-test when there is doubt about equal variance (3).

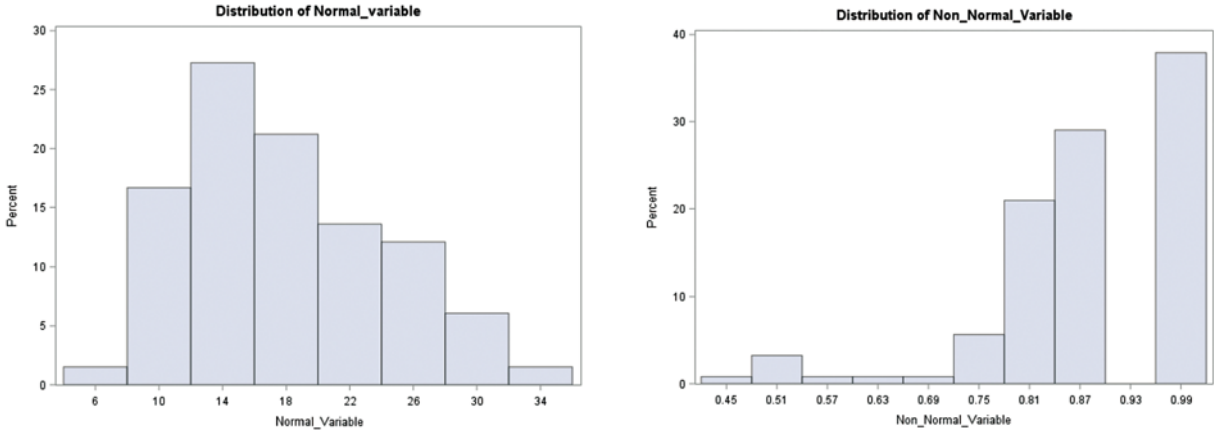

Fig. 2 Histograms of normal (or close to normal) vs. non-normal data.

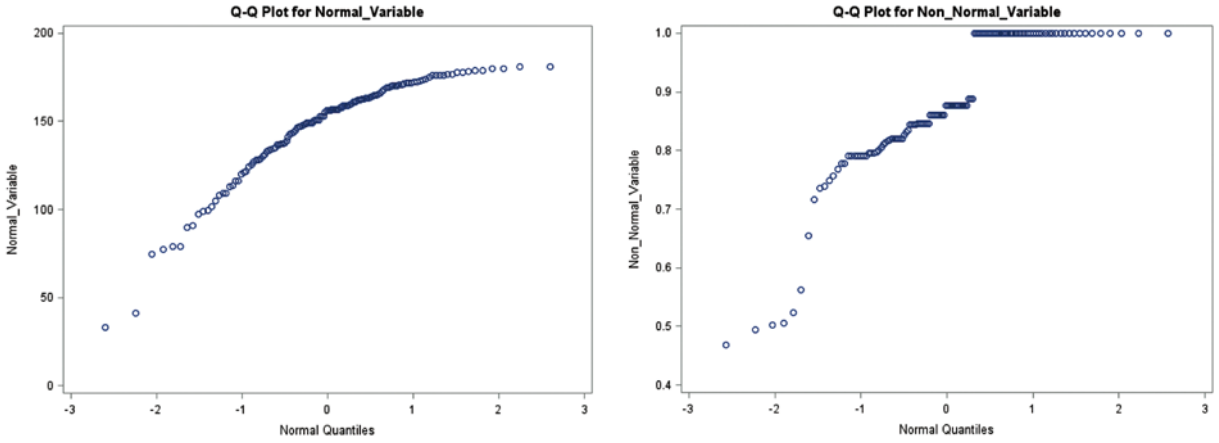

Fig. 3 Normal Q_Q plots for normal and non-normal data.

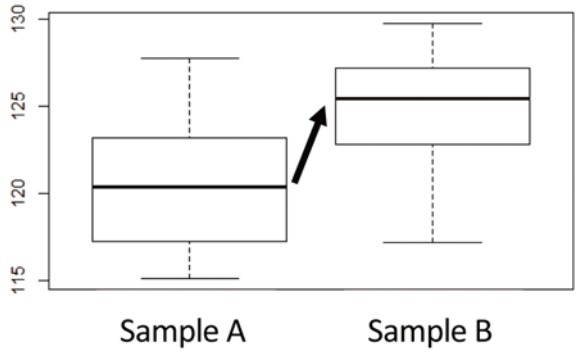

Fig. 4 Tests for shift in distribution.

such as a log transformation will convert different distribution shapes to be closer to normality but should be used with caution because of reduction in power, lack of dealing with outliers, and difficulty in interpretation of results (4).

4. Use a non-parametric procedure. Unless it is specifically critical to test for difference of means, non-parametric options to the t-test offer substantial benefits when normality of data is a concern and a small sample size is the only option.

Non-parametric options

A common non-parametric alternative to the t-test applications are the Wilcoxon non-parametric tests[2]. The Wilcoxon tests do not rely on an assumption of normality; they are based on ranks and sum of ranks of observations rather than on actual values, hence they are not affected by outliers or extreme observations. When normality is violated, the Wilcoxon tests are consistently more powerful than independent or dependent samples t tests (12, 13).

The Wilcoxon tests do not use the mean as a measure of central tendency; they are a test of shift in the distribution of responses. The null hypothesis is that median values are equal (14).

[2]There are a variety of non-parametric tests; here we focus on just the Wilcoxon tests that are comparable to t-tests. The Wilcoxon tests are identical to the Mann-Whitney U test.

A significant difference in median values suggests a shift in the overall distribution. A p-value is generated either as an exact or approximated value. Different Wilcoxon tests match the different t-test application. For

- Independent 2 sample studies: the Wilcoxon Rank Sum Test is used.
- Dependent, paired data: the Wilcoxon Sign Rank Test is used.
- Single sample test for location: the Sign Test is used.

There is also a Wilcoxon test for paired, ordinal scale data, the Sign Test, for which there is no comparable t-test.

Three assumptions are common across the four applications:

1. the study variable is, at minimum, ordinal,
2. independence of samples. For paired, dependent variables (Wilcoxon sign rank test), pairs must be randomly and independently drawn.
3. Equal variance between samples under the null hypothesis.

When the assumption of normality is met (which occurs rarely), the t-test holds a small power advantage over the Wilcoxon non-parametric tests. Estimates of the mean difference and confidence interval provide valuable summary statistics that have no equivalent in non-parametric statistics. But when normality is in question (and it often is) the mean is likely not a reliable indicator of central tendency and the Wilcoxon non-parametric tests offer substantially higher levels of power and a reliable alternative to the t-test (15) for estimating p values. If estimation of effect size is critical, in addition to estimating the p value, then an additional method may be needed for that purpose

Conclusions

t-tests are commonly mis-applied to non-normally distributed data which can affect the conclusions drawn; the null hypothesis may be rejected or fail to be rejected in error. Check data for all study variables for normality; non-normally distributed data is common. The Wilcoxon non-parametric tests for differences in distribution offer an effective alternative; when normality is violated, the Wilcoxon tests are consistently more powerful than independent or dependent samples t tests. If there is uncertainty about the normality of the data (unless estimates of mean difference and confidence intervals are important in analysis (and then consider carefully how reliable the mean is as an indicator of central tendency)), use the Wilcoxon tests instead. When in doubt, use the Wilcoxon tests.

Acknowledgments

I would like to acknowledge the contributions of my fellow biostatisticians at Boston Children's Hospital who freely gave their knowledge and critical feedback for this paper.

Sources of funding

None.

Conflicts of interests

None.

Reprint requests and correspondence:
David N. Williams PhD
Senior Biostatistician, Biostatistics and Research Design Core, Institutional Centers for Clinical and Translational Research Department of Adolescent Medicine, Boston Children's Hospital, and Instructor of Pediatrics, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115, USA
E-mail: david.williams3@childrens.harvard.edu

References

1. Hayes AF, Cai L. Further evaluating the conditional decision rule for comparing two independent means. Br J Math Stat Psychol 2007; 60: 217-44.
2. Hoekstra R, Kiers HA, Johnson A. Are assumptions of well-known statistical techniques checked, and why (not)? Front Psychol 2012; 3: 137.
3. Ruxton GD. The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. Behavioral Ecology 2006; 17: 688-90.
4. Gali S. On importance of normality assumption in using a t-test: one sample and two sample cases. International Symposium on Emerging Trends in Social Science Research Chennai, India 2015.
5. NCSS Statistical Solutions: Normality Tests 2018 [updated June, 2018. Available from: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/ Normality_Tests.pdf
6. Centre for Applied Statistics Courses, Statistics and Research Methods, Chapter 3. Summarizing Data, 2016 [cited May, 2018]. Available from: http: //www. ucl. ac. uk/ich/short-courses-events /about-stats-courses/stats-rm/Chapter_3_Content/.
7. Doane DP, Seward LE. Measuring skewness: a forgotten statistic? Journal of Statistics Education 2011; 19.
8. Ford C. Understanding Q-Q plots [Internet]. University of Virginia. 2017 [cited May, 2018]. Available from: http://data.library.virginia.edu/understanding-q-q-plots/.
9. Steiger JH. Robustness. Nashville, TN: Vanderbilt University; 2015.
10. Lumley T, Diehr P, Emerson S, et al. The importance of the normality assumption. Annu Rev Public Health 2002; 23: 151-69.
11. Grace-Martin K. Outliers: To Drop or Not to Drop, 2018, Available from https://www.theanalysisfactor.com/ outliers-to-drop-or-not-to-drop.
12. Blair RC, Higgins JJ. A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic

under various nonnormal distributions. Journal of Educational Statistics. 1980; 5: 309-34.
13. Fay MP, Porschan MA. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Statistics Surveys 2010; 4: 1-39.
14. D'Agostino RB, Sullivan LM, Beiser AS. Introductory Applied Biostatistics. Toronto: Thomson; 2006.
15. Sawilowsky SS. Misconceptions Leading to Choosing the t Test Over the Wilcoxon Mann-Whitney Test for Shift in Location Parameter. Journal of Modern Applied Statistical Methods 2005; 4: 598-600