JOURNAL OF APPLIED MATHEMATICS AND DECISION SCIENCES, 8(4), 235–246

Robustness to Non-Normality of Various Tests For the One-Sample Location Problem

MICHELLE K. MCDOUGALL† mkb@deakin.edu.au
School of Information Technology, Deakin University, Waurn Ponds, VIC 3217, Australia

GLEN D. RAYNER
Fellow, Institute of Mathematical Modelling and Computational Systems, University of Wollongong, NSW 2522, Australia

Abstract. This paper studies the effect of the normal distribution assumption on the power and size of the sign test, Wilcoxon's signed rank test and the t-test when used in one-sample location problems. Power functions for these tests under various skewness and kurtosis conditions are produced for several sample sizes from simulated data using the *g-and-k* distribution of MacGillivray and Cannon [5].

Keywords: *G-and-k* distribution, power, quantile function, sign test, size, t-test, Wilcoxon's signed rank test

1. Introduction

In the context of asymmetric data, the difference between testing for a specific mean and testing for a specified median is often critical. However, simply an idea of location is usually the object, however that may be defined. In these situations, the median is probably the most useful description of data location, but typically practitioners use mean tests as alternatives or competitors. To investigate the consequences of this point of view, we treat the Wilcoxon, t-test and sign test as tests for a specified median and analyse their properties. Of these tests, only the t-test is not already a median test.

In order to use the well known parametric t-test, the data must either have been sampled from a population that is normally distributed, or the sample size must be sufficiently large so that asymptotic normality of the sample mean can be assumed. If these assumptions cannot be made, then non-parametric procedures such as the sign test, or Wilcoxon's signed rank test should be employed. These non-parametric tests have

† Requests for reprints should be sent to Michelle K. McDougall, School of Information Technology, Deakin University, Waurn Ponds, VIC 3217, Australia.

less restrictive assumptions about the shape of the parent population than the t-test. Wilcoxon's signed rank test assumes that the sample is drawn from a continuous, symmetric population while the sign test only requires that the population is continuous around the vicinity of the median. Non-parametric tests are also usually easier to apply and understand than the corresponding parametric tests and are generally insensitive to outliers. Hollander and Wolfe [4] list nine advantages of non-parametric methods over parametric methods.

When the assumptions necessary for parametric tests are true, then parametric tests should have greater power than the corresponding non-parametric tests. This power difference may not necessarily be large however. When these assumptions are not true, there can be little confidence in the resulting inference if parametric tests are used, and in this situation non-parametric techniques will often have greater power. Carolan and Rayner [2] found that even for parametric techniques that are robust, the corresponding non-parametric tests may be superior when the parametric assumptions do not hold. In addition, Rayner and Carolan [11] state that as the data becomes increasingly non-normal, the assumption that the significance level (size) is its nominal value becomes increasingly doubtful.

Cressie [3] examined the behaviour of the t statistic when sampling from non-normal populations, using functions of the third and fourth moments about the mean as measures of skewness and kurtosis respectively. He found that heavy tailed data produced light tailed t values, and vice versa. Positively skewed data resulted in negatively skewed test statistic values and vice versa. He observed that skewness had a greater impact on the t distribution compared with kurtosis. In addition, with kurtosis only present, the size of the t test was lower than the nominal value for heavy tailed data and, for lighter tailed data, the size was greater than the nominal value.

Our interest is in testing a specified value, μ_0, of the median.

2. Quantile Distributions

The *g-and-k* distribution of MacGillivray and Cannon [5] is a quantile distribution family, that is, a distribution specified in terms of its quantile function. Details of these distributions are repeated and further explained in Rayner [9] and Rayner and MacGillivray [10]. The *g-and-k* distribution is a transformation of the standard normal distribution which introduces effects due to skewness (measures of asymmetry) and kurtosis (heaviness/lightness in the tails). This distribution includes a wide variety of shapes, as well as the normal distribution. The quantile function for the

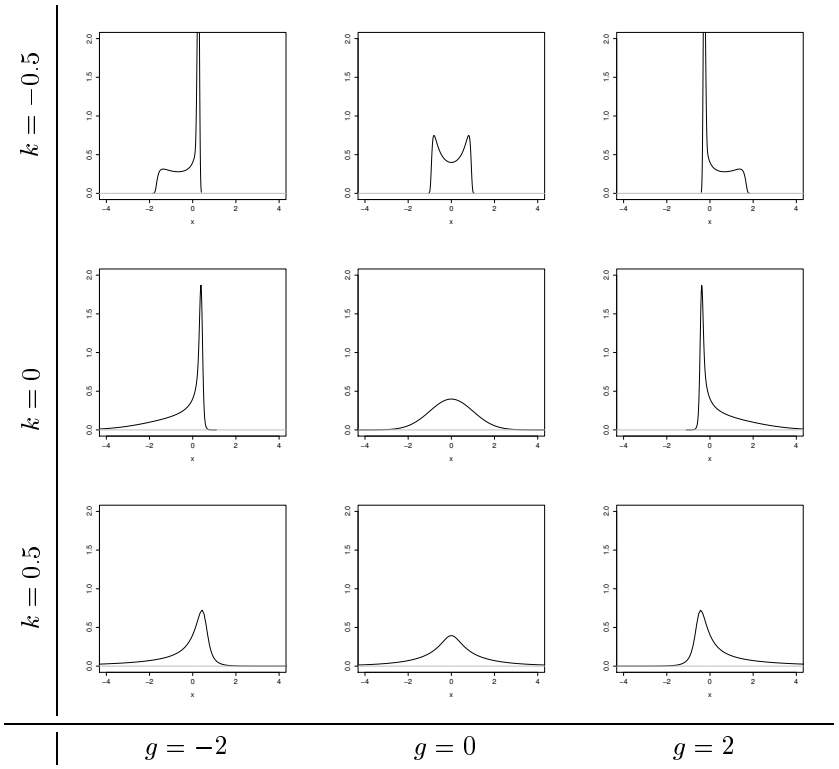

Figure 1. Graphs showing distribution shapes for g-and-k distributions with each combination of g (skewness) and k (kurtosis) used.

g-and-k distribution is:

$$\begin{aligned} Q(u|A,B,g,k) &\equiv F^{-1}(u|A,B,g,k) \\ &= A + Bz_u \left(1 + c\frac{1-e^{-gz_u}}{1+e^{-gz_u}}\right)(1+z_u^2)^k \end{aligned} \tag{1}$$

where A and $B > 0$ are location (median) and scale parameters respectively, g measures skewness in the distribution, $k \geq -\frac{1}{2}$ measures kurtosis of the distribution, $z_u = \Phi^{-1}(u)$ is the uth standard normal quantile, and c is a chosen constant. We have used $c = 0.8$ in this paper.

The sign of the skewness parameter indicates the direction of skewness: $g < 0$ indicates the distribution is skewed to the left, and $g > 0$ indicates skewness to the right. Increasing/decreasing the magnitude of g increases/decreases the skewness in the indicated direction. When $g = 0$,

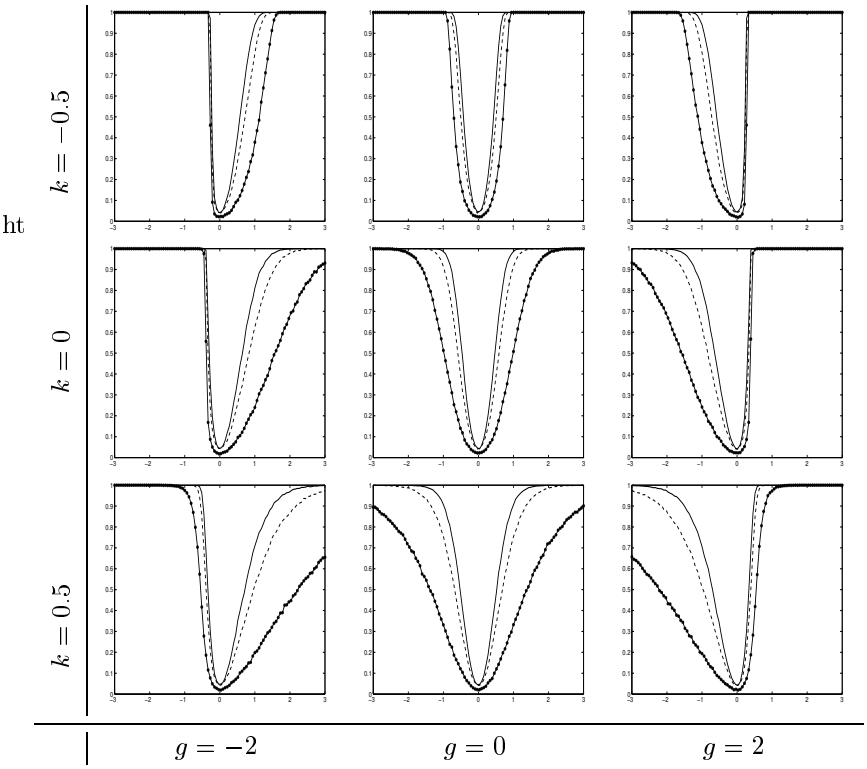

Figure 2. Power curves of the sign test, for n = 10,20,30
——— n=30
- - - - - - n=20
-*-*-*-*- n=10.

the distribution is symmetric. Increasing the kurtosis parameter, k, adds more kurtosis to the standard normal base distribution. The value $k = 0$ corresponds to no extra kurtosis added to the standard normal base distribution. When $-\frac{1}{2} \leq k < 0$, the *g-and-k* distribution exhibits less kurtosis than the normal distribution. Refer to Figure 1 for the distribution shapes of the combinations of g and k values used.

3. Tests

Let X_1,...,X_n be a random sample from a continuous population with mean μ, median M, and standard deviation σ (σ unknown). We can test hy-

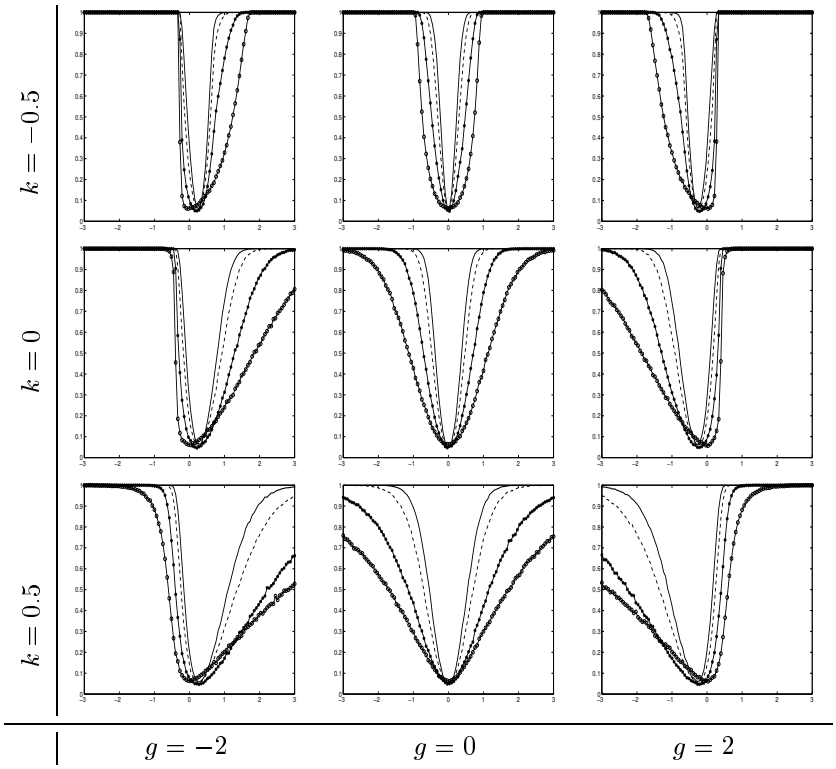

Figure 3. Power curves of the Wilcoxon signed rank test, for n = 5,10,20,30
———— n=30
- - - - - - n=20
-*-*-*-*- n=10
-o-o-o-o- n=5.

potheses about the location of this population using the well known t-test, Wilcoxon's signed rank test, or the sign test.

3.1. Test details

The t-test assumes that $X_i \sim N(\mu, \sigma^2)$ for $i = 1, ..., n$. For large samples, the assumption that the population is normally distributed is not necessary, since the sample mean is approximately normally distributed according to the central limit theorem. Wilcoxon's signed rank test assumes that the data are drawn from a continuous, symmetric population. Further details

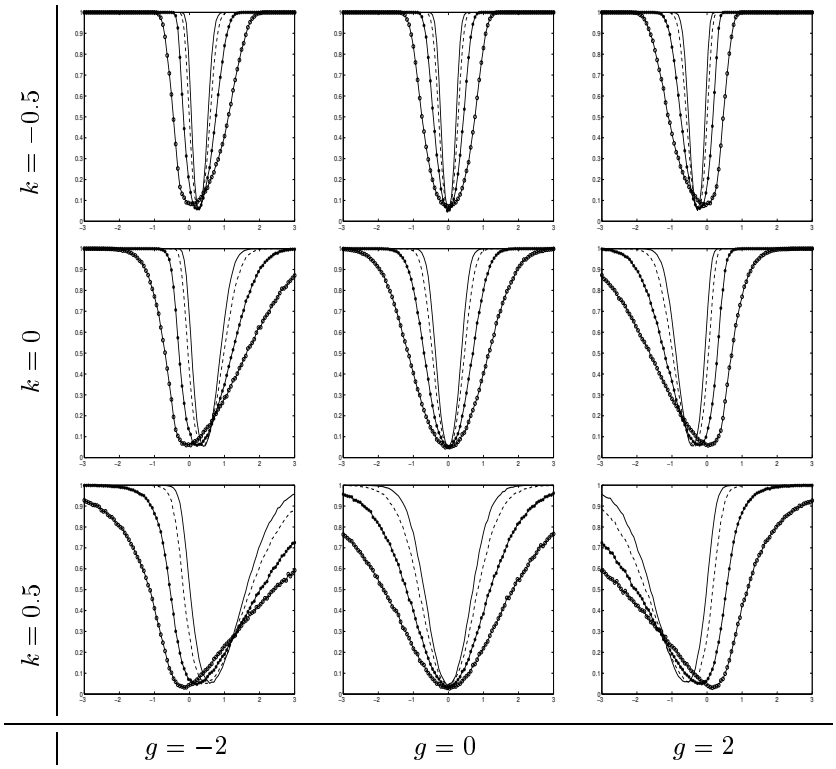

Figure 4. Power curves of the t-test, for n = 5,10,20,30
———— n=30
- - - - - - n=20
-*-*-*-*- n=10
-o-o-o-o- n=5.

about the assumptions and test statistic calculations for the t-test and Wilcoxon's test can be found in most introductory statistics texts, including Weiss [14]. There is no assumption about the shape of the population distribution for the sign test. Moore and McCabe [6] provide additional details and examples for the sign test.

3.2. Power comparison

Many statisticians assume that the t-test works reasonably well even when the data under consideration are not obviously normally distributed, and

the sample size is small or moderate, provided the data are 'not too far' from normal. That is, the t-test is widely thought to be robust to moderate violations of the normality assumption.

Rao [8] states that procedures based on t-tests (especially one sided procedures) are more sensitive to skewness than to heaviness or lightness in the tails (kurtosis). For symmetric distributions, t-test procedures perform relatively well, even when the distribution is non-normal. Ott [7] concludes that when the population distribution is symmetric but heavy-tailed, robust methods such as Wilcoxon ranked sum test are more efficient than the corresponding t-test about the parameter μ. Weiss [14] warns that t-test procedures should not be used when the data are obviously skewed.

Bradley [1] gives the asymptotic relative efficiency (ARE) of the sign test relative to the t-test when parametric conditions are met (random, independent sample from a normal population) as 0.637. When parametric conditions are not met (population non-normal), Bradley [1] finds the sign test may be more efficient than the t-test.

When parametric conditions are met, then for finite samples the relative efficiency of the sign test to the t-test increases as the values of n, α and $|\mu - \mu_0|$ decrease. Relative efficiency figures as high as 0.96 have been found and are given by Bradley [1].

Although the sign test does not assume a symmetric population, the test is clearly still valid if symmetry is assumed. However, as stated in Sprent [13], the sign test then often has lower efficiency and less power than the Wilcoxon signed rank test. Wilcoxon's test uses more information than the sign test since the magnitudes of the observations are ranked and the ranks used in the test statistic formula. If the data are skewed, the sign test often performs as well or better than the inappropriate signed-rank test. Even for some symmetric distributions (particularly those with long tails), the sign test is more efficient than the signed-rank test. For example, Sprent [13] states that for the double exponential distribution, the sign test has an asymptotic relative efficiency of 4/3 relative to the signed-rank test.

For a normally distributed variable, Weiss [14] finds that the t-test is more powerful than the Wilcoxon signed-rank test, but not much more powerful. However, if the variable is symmetric but not normally distributed, the Wilcoxon signed-rank test is usually more powerful than the t-test and is often considerably more powerful.

4. Simulation Study

We have used the MATLAB package to calculate empirical power curves, estimated using 10,000 simulations, for the sign test, Wilcoxon's signed

rank test and the t-test based on samples of sizes n= 5, 10, 20 and 30 for $\mu_0 = 0$ and at nominal size $\alpha = 5\%$. The *g-and-k* distribution (see section 2) was used to simulate data with varying kinds and amounts of non-normality. Skewness was achieved using g values of -2, 0 and 2 and kurtosis was achieved using values for k of -0.5, 0 and 0.5. Simulations were carried out with A=0 and B=1 for each of the following nine combinations of g and k, (g, k): (-2,-0.5), (-2,0), (-2,0.5), (0,-0.5), (0,0), (0,0.5), (2,-0.5), (2,0), (2,0.5). Figure 1 shows the *g-and-k* distribution shape for each combination of parameters. Note that for g=0 and k=0 the distribution is the standard normal distribution.

Using the 10,000 sets of simulated data for each n and (g, k) combination, a two-sided test for location was performed for the t-test, Wilcoxon and sign tests and the power curve empirically estimated as the proportion of simulated samples rejected. Critical values were not used; MATLAB calculated p-values for each hypothesis test and compared these with α. MATLAB finds exact p-values for the t-test, for the Wilcoxon test when $n \leq 15$, and for the sign test when $n < 100$. Otherwise a normal approximation is used for the Wilcoxon test and a continuity corrected normal approximation for the sign test.

Figures 2 to 4 show power curves in terms of the true location parameter $-3 \leq \mu \leq 3$ for various kinds of non-normal data (values of g and k).

Figure 5 shows true sizes of each test as the g and k parameters are varied over $-5 \leq g \leq 5$ and $-\frac{1}{2} \leq k \leq 1$. Shown are plots for sample sizes $n = 10, 15, 30$ chosen to illustrate the nature of the progressive separation of these size curves. Note that these plots show sizes, rather than powers. In terms of test size, this plot emphasises the adverse reaction of the t-test and Wilcoxon signed rank test to non-normality of the data, whereas the sign test is indifferent. Importantly, the effect of increasing sample size is to amplify this adverse reaction, not reduce it.

5. Results

Figures 2 - 4 show how the power curves for each of the sign, Wilcoxon and t-tests vary according to the different skewness and kurtosis combinations used for the data. Some noteworthy observations are listed below.

- All tests are more powerful for lighter tails (less kurtosis) rather than heavier tails, with the effect strongest for the sign test, followed by the Wilcoxon test, then the t-test.
- Generally all three tests are more powerful for larger sample sizes.

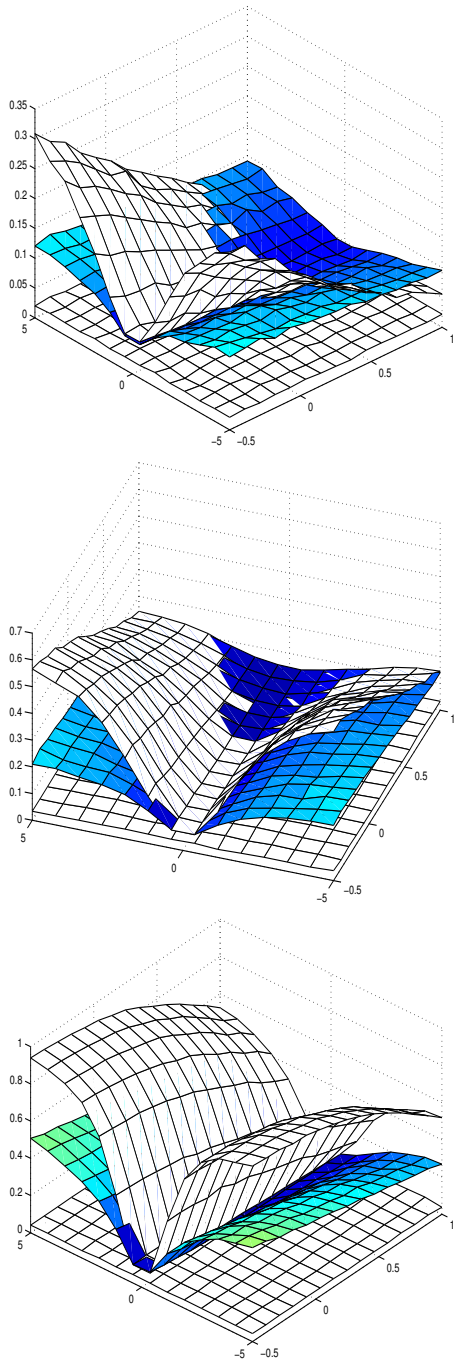

Figure 5. Comparison of sizes for the t-test (top), the Wilcoxon signed rank test (middle, shaded) and the sign test (bottom) for $-5 \leq g \leq 5$ and $-0.5 \leq k \leq 1$.

- For the smallest considered sample of $n = 5$ no sign test exists with size less than 0.05.
- Generally, skewed distributions produce tests that are more powerful at detecting location shifts in the heavier tail (presumably because there is likely to be more data there).
- For the sign test, the test size is more affected by sample size than by skewness or kurtosis.
- For the Wilcoxon and t-tests, the minimum point of the power curve seems to 'drift away' from μ_0=0 as the sample size increases. This drift depends on the amount of skewness, and moves in the opposite direction to the direction of the skewness. This means that for these tests as tests for a specified median, the true size gets worse and worse with increasing sample size if there is skewness present in the data. This is much more pronounced for the t-test than for the Wilcoxon signed rank test which is unsurprising as the t-test is an unbiased test for the mean, not the median.

By comparing curves for the three tests at the various sample sizes (graphs not included here) we make several further observations.

- All three tests are more powerful for greater sample size and flatter tailed data (less kurtosis).
- The minimum point of the power curves for the Wilcoxon test and, to a greater extent, the t-test, 'drift' from μ_0=0 quite significantly in the presence of skewness. This causes the bad empirical sizes (see Figure 5). The situation gets worse as the sample size increases, due to the assumptions being violated and this effect being more pronounced with increasing n.
- For normally distributed data and symmetric light-tailed data, there is very little difference in the power curves for the t-test and Wilcoxon's test, regardless of the sample size.

Figure 5 shows how the true size of the t-test and the Wilcoxon signed rank test is sensitive to skewness and kurtosis. Ideally the size of a location test should be insensitive to these parameters. While the sign test displays a flat power curve, both the Wilcoxon signed rank test and, to a greater extent, the t-test, display much more sensitivity to departures from normality. The effect of non-normality on these tests increases with sample size. Both the Wilcoxon and t-tests seem mainly sensitive to skewness rather than kurtosis with this sensitivity increasing for lighter tailed (less skewed) distributions.

6. Conclusion

When the normal distribution assumption holds, unsurprisingly the t-test is the most powerful test, although there is very little difference in power over Wilcoxon's test. Again, under the assumption of normality, for $n \geq 20$ the sign test has almost as much power as the t-test and Wilcoxon's test. However, the t-test and Wilcoxon's signed rank test are more sensitive to skewness than kurtosis. This result for the t test is consistent with Cressie [3].

In agreement with Ott [7], when the data are symmetric and heavy tailed, Wilcoxon's test is more powerful than the t-test. The sign test is still as powerful as Wilcoxon's test for $n \geq 20$. For symmetric, light tailed data, the t-test is more powerful (though only slightly) than Wilcoxon's test for all sample sizes. For n=5, the size of the t-test is less than the nominal value of 0.05 for heavy tailed data, and for light tailed data the size is greater than 0.05, supporting Cressie [3].

Larger sample sizes increase the power of the sign test, for any data distribution. For the t-test and Wilcoxon's signed rank test with a fixed nominal size, if the data are not symmetric, the power curves drift from a minimum at $\mu = \mu_0$, with the true size increasing with n.

For skewed data, the sign test has greater power than Wilcoxon's signed rank test or the t-test (for $n > 5$). The true size for the sign test is at most the nominal size. For the sign test, the minimum of the power curve is always at $\mu = \mu_0$.

On the basis of this study, if the data are symmetric, use Wilcoxon's signed rank test and if not, use the sign test ($n > 5$). If $n \geq 20$ the sign test can be used without significant loss in power compared with the other tests unless the distribution is light tailed.

Many statistics text books, including Weiss [14], Moore and McCabe [6] and Samuels and Witmer [12] state that the t-test is still appropriate when the data are non-normal provided n is 'large'. Suitable sample sizes given include values of the order of 15, 20, 30 or 40, depending on the degree of non-normality. However the practice of using the t-test in such situations appears to be a dangerous one. The fact that size problems for the t-test (and Wilcoxon's test) are experienced more often as the sample size increases is counter to our intuition and should serve as a warning to those relying only on the asymptotic normality of the sample mean. While this drift can to some extent be explained by the fact that the t-test detects the mean rather than the median, this is not sufficient to allow us to ignore the problem. The Wilcoxon test also suffers from this problem. Although asymptotic normality of the t-test statistic can be used to obtain

its distribution, few statisticians seem to remember that this has nothing to do with the optimality of this test.

The combinations of g and k used in this study may not all necessarily relate to situations commonly experienced by the majority of practitioners. However, in case such extreme distributions were to be encountered in practice, the behaviour of the various tests has been studied here for interest's sake as well as completeness.

References

1. J. V. Bradley. *Distribution-Free Statistical Tests*, Prentice Hall, New Jersey, p. 168, 1968.
2. A. M. Carolan and J. C. W. Rayner. One sample score tests for the location of modes of non-normal data. *Journal of Applied Mathematics and Decision Sciences*, 5(1):7–25, 2000.
3. N. Cressie. Relaxing assumptions in the one-sample t-test. *Austral. J. Statist.*, 22(2):143–153, 1980.
4. M. Hollander and D. A. Wolfe. *Nonparametric Statistical Methods*, 2nd edition, New York: Wiley, p. 1, 1999.
5. H. L. MacGillivray and W. H. Cannon. Generalizations of the g-and-h distributions and their uses. *Working Paper, 1998*
6. D. S. Moore and G. P. McCabe. *Introduction to the Practice of Statistics*, 4th edition, W. H. Freeman and Company, New York, p. 504-505, 509-511, 2002.
7. L. Ott. *An Introduction to Statistical Methods and Data Analysis*, 4th edition, PWS-Kent, p. 243 , 1988.
8. P. V. Rao. *Statistical Research Methods in the Life Sciences*, Duxbury Press, California, p. 166, 1998.
9. G. D. Rayner. *Statistical methodologies for quantile-based distributional families.* Ph.D Thesis, Queensland University of Technology (QUT), 2000.
10. G. D. Rayner and H. L. MacGillivray. Numerical maximum likelihood estimation for the g-and-k and generalised g-and-h distributions. *Statistics and Computing*, 12:57–75, 2002.
11. J. C. W. Rayner and A. Carolan. Assessing robustness of the one-sample t-test. ICOTS 5. *Proceedings of the Fifth International Conference on Teaching of Statistics*, Pereira-Mendoza, L. et al. ed., Vol. 3, pp. 1225-1232.
12. M. L. Samuels and J. A. Witmer. *Statistics for the Life Sciences*, 2th edition, p. 206. Prentice-Hall, Inc., Upper Saddle River, New Jersey, 1999.
13. P. Sprent. *Data Driven Statistical Methods*, Chapman and Hall, London, 1998.
14. N. A. Weiss. *Introductory Statistics*, 5th edition, Addison-Wesley, Reading, Massachusetts, 1999.