ASSESSING ROBUSTNESS OF THE ONE-SAMPLE t-TEST

John C.W. Rayner and A. Carolan, University of Wollongong, Australia

In statistical folklore some parametric tests are designated as generally robust, and hence almost universally applicable. Conversely some tests are supposedly so sensitive to their underlying assumptions that their use can seldom be appropriate. So the t-test is generally seen as always applicable, and Bartlett's test of homogeneity of variance is always dubious. This all or nothing approach is counter-intuitive. It is more likely that as an assumption like normality progressively fails, the assumption that the significance level is, say, 5%, progressively becomes more doubtful. The rate of decline will depend on both the test and the property in question.

We assess the one-sample t-test and the two-sample F test for equality of variance. The properties are the closeness of the actual and nominal test sizes and optimality. We give practical advice to the data analyst faced with outcomes such as those above.

INTRODUCTION

Not all users of statistical tests check assumptions such as normality and homogeneity of variance, and for those that do, the subsequent choices may be difficult. Suppose that when you apply a standard statistical package to a particular data set, you find that the one-sample t-test has p-value 0.02, the Shapiro-Wilk test assessing normality has p-value 0.05, and the Wilcoxon test has p-value 0.20. Do you assume that the one-sample t-test is robust, and conclude there is some evidence of a that the mean is other than that hypothesised, or do you doubt the validity of the t-test, and on the basis of the Wilcoxon test conclude that there is no valid evidence of against the null hypothesis? This example prompts the observations in the abstract.

For the one-sample situation we compare the t-test, the Wilcoxon test and an optimal test for the correct model. For the two-sample situation we compare the likelihood ratio F test and the nonparametric Mood test.

ASSESSING NORMALITY AND ROBUSTNESS

The parametric tests we subsequently consider are for data assumed to come from normal populations. If normality is the correct model, these tests are optimal. It is useful to assess normality and to build models of progressive failure of normality. To do this first assume we have a random sample $X_1, \ldots, X_n$ from a $N(\mu, \sigma^2)$ distribution with probability density function.

$$f_X(x;\ \mu,\ \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\ \exp\left\{-\frac{(x\ -\ \mu)^2}{2\sigma^2}\right\}, -\infty < x < \infty.$$

This can be assessed by embedding $f_X(x;\ \mu,\ \sigma)$ in a *k-parameter alternative*:

$$g_k(x;\ \theta,\ \mu,\ \sigma) = C(\theta;\ \mu,\ \sigma)\ \exp\left\{\sum_{i\ =\ 3}^{k+2}\theta_i h_i\left((x-\mu)/\sigma\right)\right\} f_X(x;\ \mu,\ \sigma),\quad -\infty < x < \infty$$

in which $\theta = (\theta_3,\ \ldots,\ \theta_{k+2})^T$ is a *k* by 1 vector of parameters, $C(\theta;\ \mu,\ \sigma)$ is a normalising constant that ensures $g_k(x;\ \theta,\ \mu,\ \sigma)$ integrates to 1 over $-\infty < x < \infty$, and the $h_r(z)$ are the normalised Hermite-Chebyshev polynomials constructed to be orthonormal on $f_X(x;\ \mu,\ \sigma)$; see Abramowitz and Stegun (1970, 22.2.15).

The $g_k(x;\ \theta,\ \mu,\ \sigma)$ were constructed by Neyman (1937) to depart smoothly from the null probability density function $f_X(x;\ \mu,\ \sigma)$. For $g_k(x;\ \theta,\ \mu,\ \sigma)$ this departure is in the moments of *X* up to the $k + 2$ th. So for example, $g_2(x;\ \theta,\ \mu,\ \sigma)$ differs from normality by having skewness and kurtosis possibly different from that of a normal distribution.

In Rayner and Best (1989, section 6.2) normality was assessed by deriving the score test of H_0^θ: $\theta = 0$ against K^θ: $\theta \neq 0$. Components V_r are defined with the property that if a particular component V_r is significantly large, it suggests the corresponding θ_r is non-zero. The score test statistic is

$$S_k = V_3^2 + \ \ldots\ + V_{k+2}^2.$$

Note that V_3 and V_4 are standardised versions of the well-known skewness and kurtosis coefficients. Rayner and Best (1989) encourage the use of S_k for formal testing, and simultaneous use of the components $V_3,\ \ldots,\ V_{k+2}$ in a data analytic manner. So a large S_4 suggests non-normality, and large V_3 and V_6 suggest deviations from what might be expected under normality in the third and sixth moments. But note the discussion in Rayner, Best and Mathews (1995).

This assessment of normality suggests an alternative model if normality is rejected. It follows that if only V_3 and V_6 are significantly large, a model of the form $g_k(x;\ \theta,\ \mu,\ \sigma)$ is suggested, but containing only θ_3 and θ_6. Inference about μ could then be based on this, the "correct" model. Alternatively a nonparametric test could

be used, or we could proceed with the parametric t-test in spite of the failure of the model. The justification for doing this is that the parametric test is "robust", but what does robust mean?

A procedure is said to be robust if its behaviour is relatively insensitive to slight departures from the assumptions underlying that procedure. There are two types of robustness of interest here. Size (sometimes "validity") robustness occurs when the nominal and actual test sizes are not significantly different under slight model failure. Optimality (sometimes "efficiency") robustness occurs when a specified optimality is not significantly affected by slight model failure. Clearly other notions of optimality, for example, independence robustness, could be defined and are of obvious interest.

Subsequently we assess size and optimality robustness by simulation studies of two situations that assume normality. Normality is allowed to progressively fail by sampling from a distribution of the form $g_k(x;\ \theta,\ \mu,\ \sigma)$ in which θ starts with a value of zero and increases in magnitude.

ONE-SAMPLE LOCATION TESTS

Suppose we assume we have a random sample $X_1,\ \ldots,\ X_n$ from a $N(\mu, \sigma^2)$ distribution, and we wish to test $H_0\colon \mu = \mu_0$ against $K\colon \mu \neq \mu_0$. The two-sided one-sample t-test rejects the null hypothesis $H_0\colon \mu = \mu_0$ in favour of $K\colon \mu \neq \mu_0$ when T^2 is sufficiently large, where $T = (\overline{X} - \mu_0)\sqrt{n} / S$ in which $\overline{X}$ and S are the mean and standard deviation of the X_i.

To examine the effect of progressive model failure, we conducted a simulation study in which observations were assumed to come from a random sample of size $n = 50$ from a distribution with probability density function

$$g(x;\ \theta_4,\ \mu,\ \sigma) = \mathrm{C}(\theta_4;\ \mu,\ \sigma)\ \exp\left\{\theta_4 h_4\left((x-\mu)/\sigma\right)\right\} f_X(x;\ \mu,\ \sigma),\ -\infty < x < \infty.$$

Probability density functions of this form for varying θ_4 are given in Figure 1. A motivation for this family would be that S_4 has been applied and found to be significant, with V_4 significantly large and with V_3, V_5 and V_6 not significantly large.

We test H_0 against K, with $\mu = 0$ and σ a nuisance parameter, without loss of generality set equal to 1 in the simulations. The nominal size was 5%. Sizes and powers were simulated for various θ_4 using 5,000 simulations.

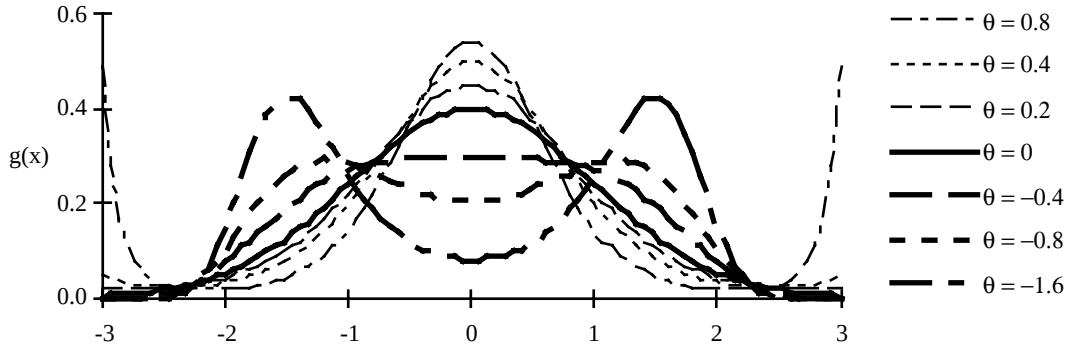

Figure 1. The probability distribution function $g(x;\ \theta_4,\ 0,\ 1)$ for varying values of θ_4

The t-test is known to be uniformly most powerful unbiased level α (see Lehmann, 1959) if the data are normal, but when $\theta_4 \neq 0$ this is not the case. For non zero θ_4 we derived the score test. This test is quite complicated, and no details are given here, other than the graphically presented simulated sizes and powers.

Note that the probability density function $g(x;\ \theta_4,\ \mu,\ \sigma)$ is symmetric, so the Wilcoxon test that tests if the median is zero, is a competitor for the t-test that tests if the mean is zero, since here the median and the mean are both zero.

Our simulations show that for all q_4 the test sizes are comparable, that when $\theta_4 = 0$ there is no real difference between the power curves, and that as θ_4 increases the score test quickly becomes dominant. The Wilcoxon test is inferior to the t-test for $-1.2 < \theta_4 < 0$ and thereafter becomes superior. The results are most effectively shown graphically. See Figure 2.

If all the score test power functions are graphed together, it is seen that as θ_4 increases the power for a given μ increases. The reverse is true for the t-test, while there is no clear pattern for the Wilcoxon test. See Figure 3.

It seems that in terms of comparability of nominal and actual test sizes, for the models considered here all tests are size robust. The huge power loses of the t-test relative to the score test shows that in terms of retaining near optimality, the t-test is not optimality robust.

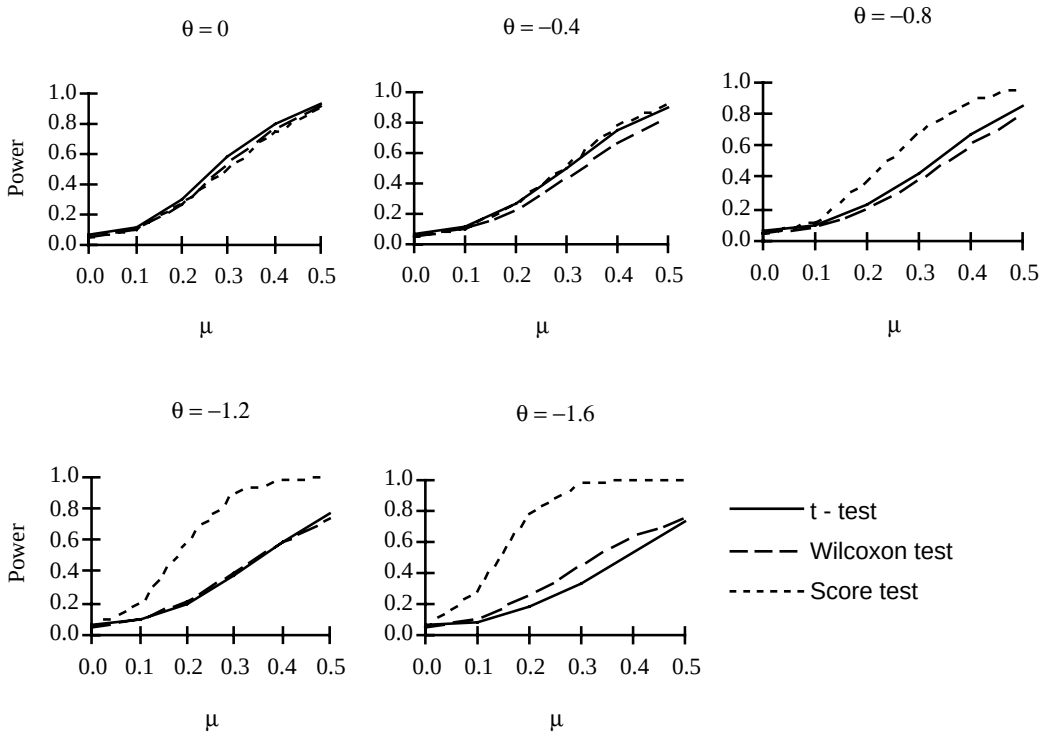

Figure 2. Comparison of power curves for testing H_0: $\mu = 0$ against K: $\mu \neq 0$ using the t-test, Wilcoxon test and score test as data becomes progressively more non normal. Based on 5000 simulations, $x \sim g(x;\ \theta_4,\ \mu,\ 1)$, $\alpha = 0.05$.

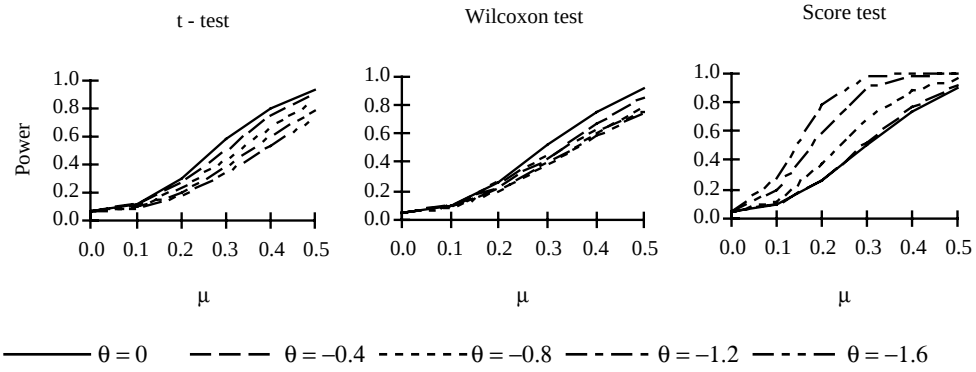

Figure 3. Power curves, as in Figure 2, grouped together for each of the three tests.

TWO-SAMPLE TESTS FOR EQUALITY OF VARIANCE

Our perception is that the t-test is thought to be generally robust, when in fact the study of the previous section shows it does not have optimality robustness. We now consider two two-sample tests for equality of variances. The F test, based on a simple multiple of the quotient of the sample variances, is the likelihood-ratio test under the assumption of normality. This test is compared with Mood's test: see for example Daniel (1990). We show that the F test has neither size nor optimality robustness.

Observations were assumed to come from two random samples, both with size $n = 50$, the first having standard deviation σ_X and the second having standard deviation σ_Y. Both are from distributions with probability density function $g(x;\ \theta_4,\ \mu,\ \sigma)$ given above. Using both the F and Mood tests, we test H_0: $\sigma_X = \sigma_Y$ against K: $\sigma_X \neq \sigma_Y$, where without loss of generality we set $\sigma_X = 1$. Again the test size is nominally 5% and sizes and powers were simulated for various θ_4 using 5,000 simulations. In this case we have not derived the score test.

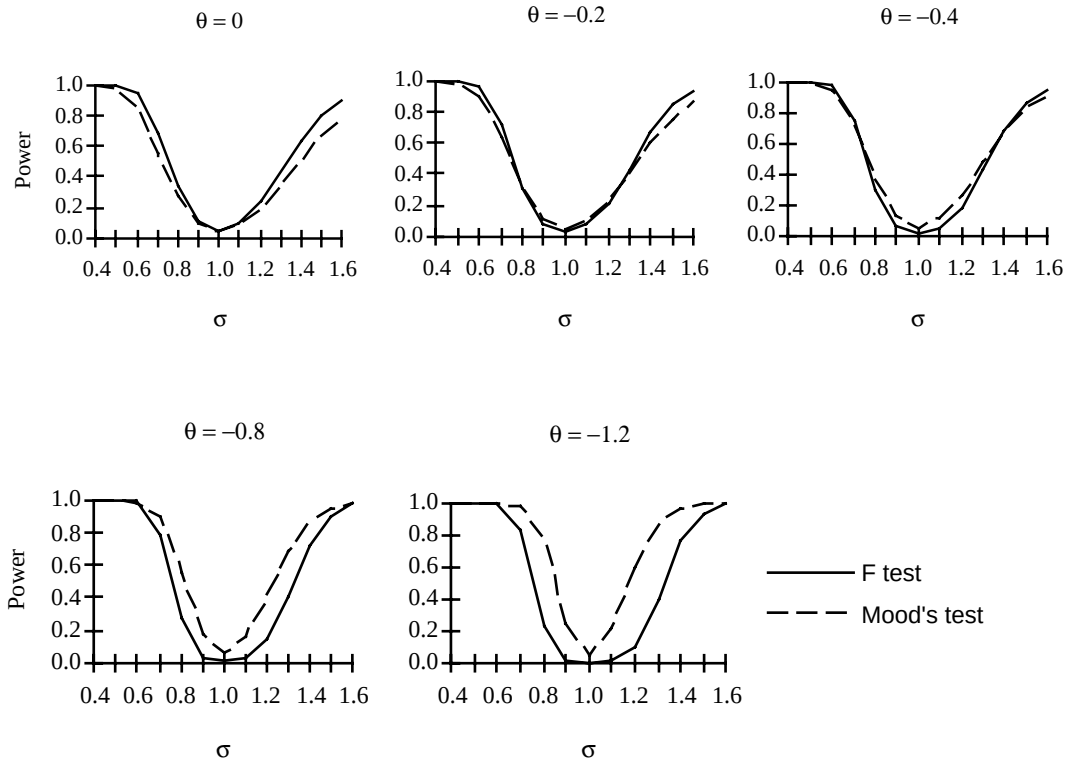

Figure 4. Comparison of power curves for testing H_0: $\sigma_X = \sigma_Y$ against K: $\sigma_X \neq \sigma_Y$ using the F test and Mood's test as data becomes progressively more non normal. Based on 5000 simulations, $x \sim g(x;\ \theta_4,\ 0,\ \sigma_x = 1)$ and $y \sim g(y;\ \theta_4,\ 0,\ \sigma_Y = \sigma),\ \alpha = 0.05$.

For normal data the F test was more powerful, while for $\theta_4 = -0.2$ and -0.4 there was little to choose between the two tests. For larger θ_4 the Mood test was significantly more powerful. This was due, at least in part, to the F test having test size significantly less than the nominal significance level, while the Mood test size was comparable with the nominal significance level. See Figure 4.

If all the Mood test power functions are graphed together, it is seen that as θ_4 increases the power for a given μ increases. There is no clear pattern for the F test. The

F test exhibits neither size nor optimality robustness. See Figure 5. Comparing the location testing of the previous section with the dispersion testing of this section, it is true that the parametric test here breaks down for smaller θ_4 than for the location tests. In this sense the t-test is more size robust than the F test.

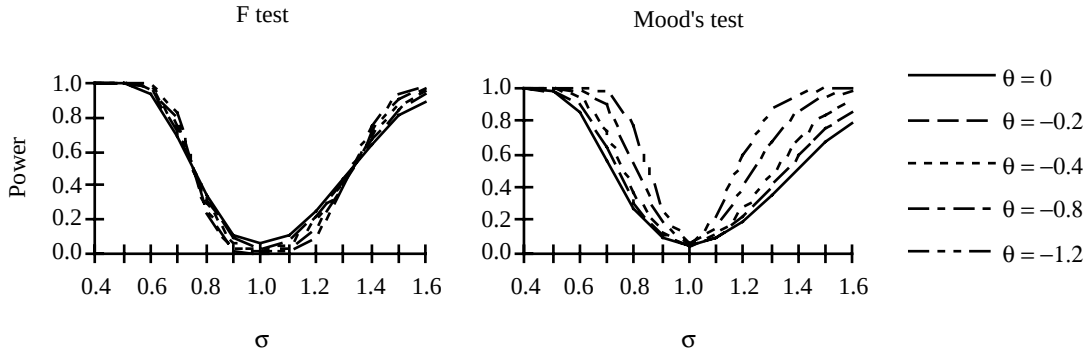

Figure 5. Power curves, as in Figure 4, grouped together for each of the tests.

CONCLUSIONS

- There are several possible notions of robustness. Teachers please note!
- In the definitions of robustness "slight" model failure is rarely defined. It may be helpful to data analysts to develop recommendations in terms of, for example, the kurtosis test being significant at the 5% but not the 1% level.
- It is unreasonable to expect any test to be universally robust.
- For tests that are not size robust, resampling p-values would be more useful than model-based possibly asymptotic p-values.
- When the model fails, tests based on the admitedly quite complicated "correct" model may have far more power than the standard tests, whether parametric or not.

REFERENCES

Abramowitz, M. and Stegun, I. A. (1972). Handbook of Mathematical Functions. New York: Dover.

Daniel, W. (1990). Applied Nonparametric Statistics (2nd ed.). Boston: PWS Kent.

Lehmann, E. L. (1959). Testing Statistical Hypotheses. New York: John Wiley.

Neyman, J. (1937). 'Smooth' test for goodness of fit. *Skand. Aktuarietidskr., 20,* 150-199.

Rayner, J. C. W. and Best, D. J. (1989). Smooth Tests of Goodness of Fit. New York: Oxford University Press.

Rayner, J. C. W., Best, D. J. and Mathews, K. L. (1995). Interpreting the skewness coefficient. *Commun. Statist.-Theor. Meth., 24(3),* 593-600.