Normal Enough? Tools to Aid Decision Making

Neil H. Spencer*, Margaret Lay, Lindsey Kevan de Lopez

Statistical Services and Consultancy Unit, Hertfordshire Business School, University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, UK

When undertaking quantitative hypothesis testing, social researchers need to decide whether the data with which they are working is suitable for parametric analyses to be used. When considering the relevant assumptions they can examine graphs and summary statistics but the decision making process is subjective and must also take into account the robustness of the proposed tests to deviations from the assumptions. We review the contemporary advice on this issue available to researchers and look back to the roots of hypothesis testing and associated work undertaken by eminent statisticians since the 1930s. From this we create a set of flow charts to give researchers tools they can use to make decisions in a more objective manner.

Keywords: quantitative data analysis; assumptions; hypothesis testing; normality; robustness

Introduction

It is well known that when undertaking hypothesis testing, parametric tests have more power than non-parametric tests providing the distributional assumptions behind the parametric tests are met. Whilst it is sometimes possible to rely on existing knowledge of the distribution of a variable in the population as a whole, it is frequently the case that judgements about the distribution (typically whether it is *sufficiently Normal*) are based on an examination of the sample data available at the time of analysis. There are a range of graphical and numerical summaries that can be created to assist in making a decision about whether the distributional assumptions are satisfied but, as will be shown below, there are no definitive guidelines and only vague, inconsistent and sometimes unhelpful rules of thumb. The researcher is thus left to exercise his or her professional judgement in an ad hoc manner. This paper aims to rectify this state of affairs by examining past and present literature on the subject and producing a set of flow charts which can be used to aid the decision making process when having to judge whether or not it is appropriate to undertake parametric analyses.

We concentrate on the most commonly used univariate tests of hypotheses for scale/interval data. Data of this type have long been analysed by social scientists and their prevalence has increased over recent years (e.g. the increase in prominence of biosocial research and the automatic gathering of data from people's digital lives). It is thus vital that social scientists are able to make appropriate decisions as to how to undertake comparisons of location for one sample (or paired samples), comparisons of location for two samples or comparisons of location for more than two samples (the analyses considered in this paper).

Below we look at the advice on this matter given in contemporary textbooks before turning to the historical roots of hypothesis testing. We briefly consider non-parametric testing and give details of the creation of flow charts to aid the decision making process before concluding with a discussion.

Guidance in contemporary statistical literature

A range of contemporary statistical and quantitative methods textbooks were selected for review (published since 2000 and, for reasons of practicality, written in English). Textbooks

* N.H.Spencer@herts.ac.uk

were chosen as these are the primary sources that are used when statisticians and researchers are learning about hypothesis testing and are thus the most influential. As well as general textbooks on statistics, those targeting specific disciplines were reviewed (social sciences but also psychology, business/management, engineering, nursing and healthcare, medicine). Clearly, given the multitude of such books, it was not possible to review every publication that met these criteria but efforts were made to include books which are widely used in the research communities.

Space precludes a comprehensive description of the results of this review but we have identified a number of themes which emerge and these are described below. Books varied in their treatment of the issue depending on the authors' approaches to the analysis of data, some preferring to emphasise detail and precision of the scientific method, some taking a pragmatic approach, seeing statistical methods simply as a tool to be used to reach one's goal, and others falling between these positions. There is no single correct approach and it is for this very reason that we make the contribution found in this paper.

In general, however, we found that guidance often lacked consistency and specificity whereby vague terms such as "modest", "large" and "extreme" are often used. This is illustrated by the following example from a book where the authors aim to emphasise a correct general approach to hypothesis testing rather than the mathematical specifics.

> "The t test is robust to moderate violation of its assumptions." … "The t test is not robust with respect to the between-samples of within-samples independence assumptions, nor is it robust with respect to extreme violations of the normality assumption unless the sample sizes are extremely large."
> (Grove, Burns & Gray 2013, p. 581)

Without clear guidance on what may be considered "extreme" violations the reader is no further forward in their quest to assess whether their data is sufficiently Normal for parametric tests to be used. However, one must not be too critical of the authors because they are faced with giving their readers the best advice available and in the absence of any definitive rules on this matter, they are presenting accepted wisdom even if it is limited.

Flow charts are an excellent way of presenting guidance to navigate complex situations and several of the statistical textbooks we reviewed included flow charts for selecting tests appropriate to the research question being asked (e.g. Salkind, 2014; Field, 2013). However, none of the books we reviewed had a flow chart or other device which helped to decide whether a distribution was sufficiently Normal for parametric tests to be used.

Advice to use histograms and plots

The most commonly mentioned advice given in the sources we reviewed was to create a histogram and if it looks "fairly" or "approximately" Normal then parametric testing can be implemented. This advice is again non-specific and relies on subjective opinion as to the relative Normality of the shape of the histogram. An experienced statistician may regard a particular histogram as Normal enough, whilst a non-statistician with less experience in analysing data may believe that it deviates too much from Normality for parametric tests to be confidently applied.

Advice to use tests of Normality

The use of statistical tests, such as the Kolmogorov-Smirnov and Shapiro-Wilks tests to check if data are Normally distributed is sometimes advised. However, it is well known that these tests are of limited use because it is easy to obtain statistically significant results when there are only small deviations from Normality if the sample size is sufficiently large. Field

(2013). whilst suggesting these tests as possible ways of identifying non-Normality, does sound a warning.

> "If you insist on using them ... always plot your data as well and try to make an informed decision about the extent of non-normality based on converging evidence."
> (Field 2013, p. 185)

Advice to use parameters of sample distribution

Some guidance recommends measures of both skewness and kurtosis for testing Normality. However, we found a lack of clear guidance across the reviewed texts relating to degree of skewness. The example given below comes from a book that gives due consideration to historical literature discussed elsewhere in this paper, but even this author has to resort to the ambiguous word "markedly".

> "...this level of accuracy is not intolerable. The same kind of statement applies to violations of the assumption of normality, provided that the true populations are roughly the same shape or else both are symmetric. If the distributions are markedly skewed (especially in the opposite directions) serious problems arise unless their variances are fairly equal."
> (Howell, 2002; p.215)

De Vaus (2002, p. 76) addresses a number of "problems" in data analysis including the issue discussed here. One of several approaches he suggests is using a "rule of thumb" that if skewness is greater than 1.0 the distribution is non-symmetrical. Abu-Bader (2010) encourages "careful data inspection and evaluation to ensure that certain conditions are met" (p2), suggesting the standardization of skewness and kurtosis and comparing the resulting scores against a normal distribution.

Some authors advised using combinations of parameters to help assess whether a distribution could be considered "Normal enough". For example, Fowler et al. (2002) takes a pragmatic approach and suggests that readers make a decision about Normality using the following suggestion.

> "… plot out the data and see if they look normal. As a back- up, calculate the mean and standard deviation of the sample and see if about 70% of the observations fall within the interval $\bar{x} \pm s$."
> (Fowler et al. 2002, p.89).

Advice to take group sizes into account

Equality of the sample sizes in each group is a factor for consideration in the question of whether a distribution is sufficiently Normal when there are two or more groups or samples. Field (2013) takes a more thorough approach than others, referring to previous literature on the subject (Donaldson, 1968; Glass, Peckham and Sanders, 1972; Lunney, 1970; Wilcox, 2012) as providing some evidence to suggest the following.

> "… when group sizes are equal the F-statistic can be quite robust to violations of Normality."
> (Field 2013, p. 444)

Although such advice may be reassuring to some, others may rightly question the meaning of "quite robust".

Advice to sake sample size into account

Some books we reviewed advised that if the sample is "large enough" there is no need to worry about violating the assumption of Normality due to the central limit theorem. Field (2013) stated that the widely accepted value for how large a sample has to be for this to apply was 30. This advice however was not consistent across sources we reviewed. For example, de Vaus (2002, p. 79) suggests that a sample size of 100 or more would make it "reasonable to use statistics that assume a normal distribution".

Advice to not worry about it

Some authors play down the importance of the Normality assumption. For example, Salkind (2014) seeks to play down the complexity of the mathematical issues, placing discussion of assumptions in a "Tech Talk" section. For the two-sample t test and the assumption of equal variances, he states the following.

> "Don't knock yourself out worrying about these assumptions because they are beyond the scope of this book."
> (Salkind 2014, p. 202)

Advice to use non-parametric tests

Some other texts advise that readers use non-parametric tests when the assumption of Normality is violated without suggesting how to judge whether it is Normal enough. For example, although Swift and Piff (2010) are very precise when it comes to the details of carrying out hypothesis tests, they state the following.

> "Most confidence intervals or tests are based on some assumptions; for instance, that the population is normally distributed or that the variance of two populations is the same. You should always be aware of these and consider their suitability for the data before you use the results of the confidence intervals or test.
> In Chapter S5 we introduce non-parametric tests. These make fewer assumptions and so can be applied more widely than the tests we have used so far."
> (Swift and Piff 2010, p. 564)

Approaches and solutions to the issue of Normality since 1930

Since the 1930s, many papers have been published concerning the issue of non-Normality and the impact this has on different tests. There is insufficient space here to give a comprehensive review of these papers but in order to provide an adequate illustration of developments that have happened, we will give a brief chronology using a selected statistical text from each decade.

In the 1930s, between the publication of the first editions of Fisher's "*Statistical Methods for Research Workers*" (1925) and "*The Design of Experiments*" (1935) from which roots much of modern-day hypothesis testing come, Egon Pearson published a paper (1931) which addressed the issue of non-Normal variation in ANOVA and what was by now called two-sample t-tests. In this Fisher looked at the effect of skewness and kurtosis as departures from Normality.

Moving to the 1940s, work on identifying the effects of non-Normality was continued by Geary (1947). The thrust of his message can be seen by his exhortation to print the following phrase in all statistical textbooks:

> "Normality is a myth; there never was, and never will be, a Normal distribution" although he goes on to say, "This is an over-statement from the practical point of view, but it represents a safer initial mental attitude than any in fashion during the past two decades"
> (Geary 1947, p. 241)

Thus, in the 1930s and 1940s, the focus was upon identifying conditions when Normality was and was not a problem for the analyses being undertaken. Where non-Normality might be problematic, adjustments to the calculation of p-values were identified as being a solution. However, this changed to a large degree in the 1950s with the development and growth in popularity of non-parametric tests. Nevertheless, it was recognized by authors such as Box (1953) that under certain conditions, two-sample t-tests and ANOVA were quite robust to departures from Normality.

During the 1960s (Boneau, 1960) examined the performance of parametric tests when the Normality assumption is violated and addressed the issue of comparative power between these and non-parametric tests. It could be argued that there has been little advancement in the subject since the 1960s.

In the 1970s a review paper by Glass et al (1972) provides an excellent, detailed history of the issues. However, it does not give concrete guidance to researchers on deciding on what to do with their analyses. The 1980s and 1990s saw the publication of papers (e.g. Blair and Higgins, 1985; Markowski and Markowski, 1990) that re-emphasized findings outlined in Boneau (1960). This concerned the potential for non-parametric tests to be more powerful than parametric tests in the presence of non-Normality. Moving to the current century, Khan and Rayner (2003) and Lantz (2013) also demonstrate the contrasting effects of non-Normality on the performance of parametric and non-parametric tests, but now using simulated data.

Drawing on the papers mentioned above and others shown in the references, we form some general conclusions as follows. Detailed conclusions relating to specific tests are presented later.

- Researchers are worried about violating assumptions (e.g. Boneau, 1960; Glass et al, 1972).
- Adjustments for non-Normality can be performed by transforming data or adjusting tests (e.g. Pearson, 1931).
- Testing for non-Normality is not straightforward (Cochran, 1947).
- Research into this issue has not advanced much in recent decades (evidence described above).

Non-parametric tests

Although non-parametric tests are mentioned above, it is appropriate to pause here and mention a common misunderstanding. Developed as a means of analysing ordinal data, their more frequent use in statistical analyses began in the 1950s. They also gained popularity due to the perception that they allow researchers analysing continuous data to not be concerned about departures from Normality. Indeed many researchers have been incorrectly led to believe that non- parametric tests require no assumptions at all to be made about the data when in fact independence is required and also equality of variation for non-parametric equivalents of ANOVA or two-sample t-tests (Zimmerman, 1998).

Consideration of Tests

Comparison of location for one sample or paired samples

We consider the univariate paired samples test as a special case of the univariate one sample test with the difference between the pairs creating the one sample. For the parametric one-sample t-test, it is assumed that the sample cases are independent of each other but also that the population from which the sample data come has a Normal distribution. We consider departures from Normality in terms of skewness and kurtosis and draw upon the work of Geary (1947) and Barrett & Goldsmith (1976). Other departures such as bimodality are indicative of issues with the sample/population beyond its distribution and are not addressed here.

Geary's work demonstrates that departures from Normality in terms of kurtosis do not affect the performance of the one-sample t-test. However, for a skewness value of 0.5, the t-test is compromised when a two-tailed test is being carried out but not for a one-tailed test. With larger values of skewness (the reciprocal of root 2 and above), the test is compromised for both one and two-tailed situations. In these circumstances, Barrett & Goldsmith (1976) show that a sample size of 40 or more is sufficient to overcome this problem.

We translate the findings of Geary (1947) and Barrett & Goldsmith (1976) into practical guidance for researchers into a flow chart shown in Figure 1. At two positions in the flow chart, a decision is made based on the value of the skewness. At one point the cut-off point has been set at ±1/2. This comes from Geary's finding that for this level and below, the one-tailed t-test is not compromised whereas at $\pm 1/\sqrt{2}$ it is compromised. A conservative approach has thus been taken to direct the user away from the t-test at skewness values further from zero than ±1/2. At the other position in the flow chart, the cut-off has been made at ±1/4. For a two-tailed t-test, Geary has shown that a skewness of 0 causes no problems but by the time at ±1/2 has been reached, the test is compromised. A pragmatic decision has thus been made to draw the cut-off at ±1/4 which represents a value balancing the need to be some distance from both the known potential for problems at ±1/2 and the implausibility of a skewness of exactly zero.

It could be argued that further analyses along the lines of Geary (1947) could be conducted to examine the performance of various cut-offs between 0 and ±1/2. However, to this we pose the counter-argument that there will be no point in this range at which the test suddenly goes from being acceptable to being compromised and any search for such a point is futile. We also argue that the notion of searching for precise values is contrary to the true nature of hypothesis testing which must take into account notions such as sampling variation and arbitrarily used levels of significance.

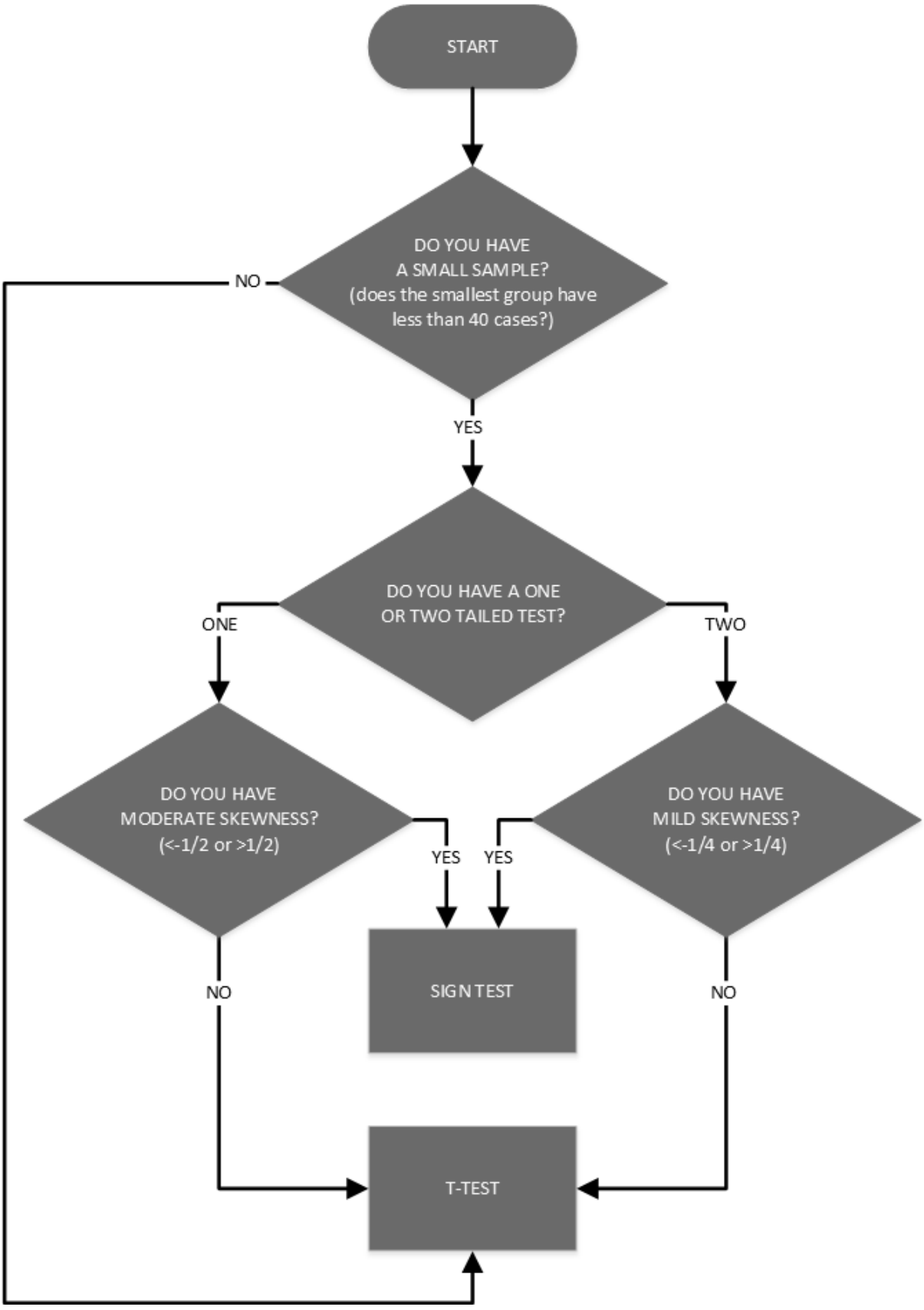

Figure 1: Flow Chart for One-Sample or Paired-Samples Test

It should be noted that the non-parametric alternative to the t-test suggested in Figure 1 is the Sign test rather than the Wilcoxon signed-rank test. As the latter makes an assumption that the data is symmetric, it is not a suitable alternative to the t-test when data are skewed.

Comparison of location for two samples

Compared with testing the location of one sample, in addition to the distributional characteristics (skewness, kurtosis) of an additional sample and the number of tails being considered, there are additional aspects to take into account for testing the location of two samples, namely equality of variation and equality of sample sizes.

The literature concerning the performance of the two-sample t-test under non-Normality is considerably larger than that for the one-sample test and, in turn, the pooled-variances version has received greater attention than Welch's version for separate variances. Conclusions that can be drawn from an examination of this literature are summarized below.

- There is a general consensus that two-tailed tests are not particularly sensitive to non-Normality (Box, 1953; Gayen, 1950b; Pearson, 1931).
- Two-tailed two-sample t-tests are less sensitive to Normality than one-tailed versions (Cochran, 1947; Tukey, 1948; Gayen, 1950b; Boneau, 1960).
- The pooled t-test is robust to heterogeneity of variances if sample sizes are equal but not otherwise (Boneau, 1960; Markowski and Markowski, 1990).
- The two-sample t-test is more robust to inequality of variances and differential skewness if sample sizes are equal (Bartlett, 1935; Gayen, 1950b; Boneau, 1960).
- Skewness can have an effect but it is diminished if sample sizes are equal or the skewness is the same in both groups (providing we are considering a two-tailed test) (Geary, 1947; Boneau, 1960; Glass et al, 1972; Wilcox, 2012).
- Kurtosis has only a minor effect, particularly if sample sizes are similar (Bartlett, 1935; Gayen, 1950b).
- Departures from Normality in terms of skewness and kurtosis are less important if sample sizes are sufficiently large (Boneau, 1960; Glass et al, 1972).
- A sample size of 80 will, for practical purposes, remove the effect of extreme skewness (Ratcliffe, 1968).

We translate these findings into practical guidance for researchers into flow charts shown in Figure 2 and Figure 3. In these flow charts, decisions are made at various stages based on how equal the sample sizes are; how large they are; how similar the skewness values are for the two groups; how equal the variances are; the skewness values and whether the test is one or two-tailed. Rationale for any cut-offs used in these decisions follow below.

For the decision as to whether the sample sizes are nearly the same, we make use of the work of Pearson and Adyanthāya (1929) and Box (1953). These authors consider sample size to be different if one is at least twice the other, so here we take a conservative approach and consider them to be different if one sample size is more than 50% larger than the other.

When it comes to deciding whether or not the skewness of the two groups is sufficiently similar, we can look at Geary (1947) and Lindquist (1953) who use $1/2$, $1/\sqrt{2}$, 1 and 1.4 to demonstrate the effect of a range of skewness values. These values are increasing by a factor of approximately 40% each time. Hence, for the purposes of the decision making here, we consider an increase of less than 20% (equivalently a decrease of 16.667%) to represent similar skewness values (subject to them both being positive or negative). We also consider that if both groups have skewness less than $\pm 1/4$ then they are similar.

The decision as to whether the variances are sufficiently similar is informed by the work of (Box, 1954). When examining different ratios for the variances he did not consider

anything less than a multiplier of three to be worth investigating. Harwell et al (1992) considered Box's three to one ratio to be modest. Taking a conservative approach here, we consider variances to be different if one is twice the size of the other or more and mildly different if one is 50% larger than the other or more.

To decide whether skewness values are moderate, large or extreme, we look to Geary (1947) and Box (1953) and use cut-offs of $\pm 1/2$, $\pm 1/\sqrt{2}$ and ± 1 respectively.

In deciding that the smallest group should have no fewer than 15 cases (so the overall sample size is at least 30) to overcome all but extreme skewness, we rely on the work of Boneau (1960). To deal with extreme skewness, Ratcliffe (1968) suggests that at least 40 cases per group are required but this extreme skewness could be an indicator that other issues also need to be addressed.

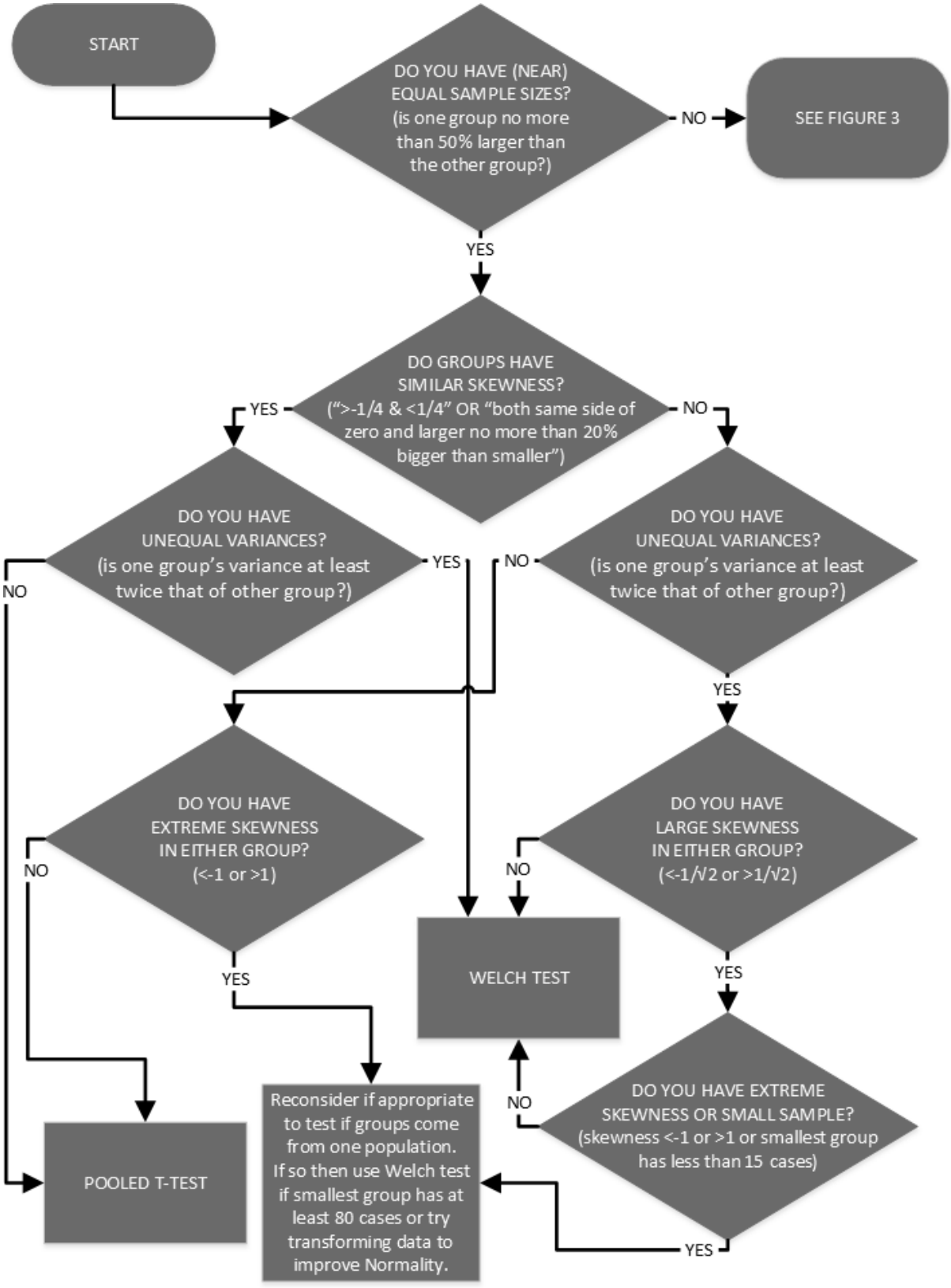

Figure 2: Flow Chart for Two-Sample Test (part 1)

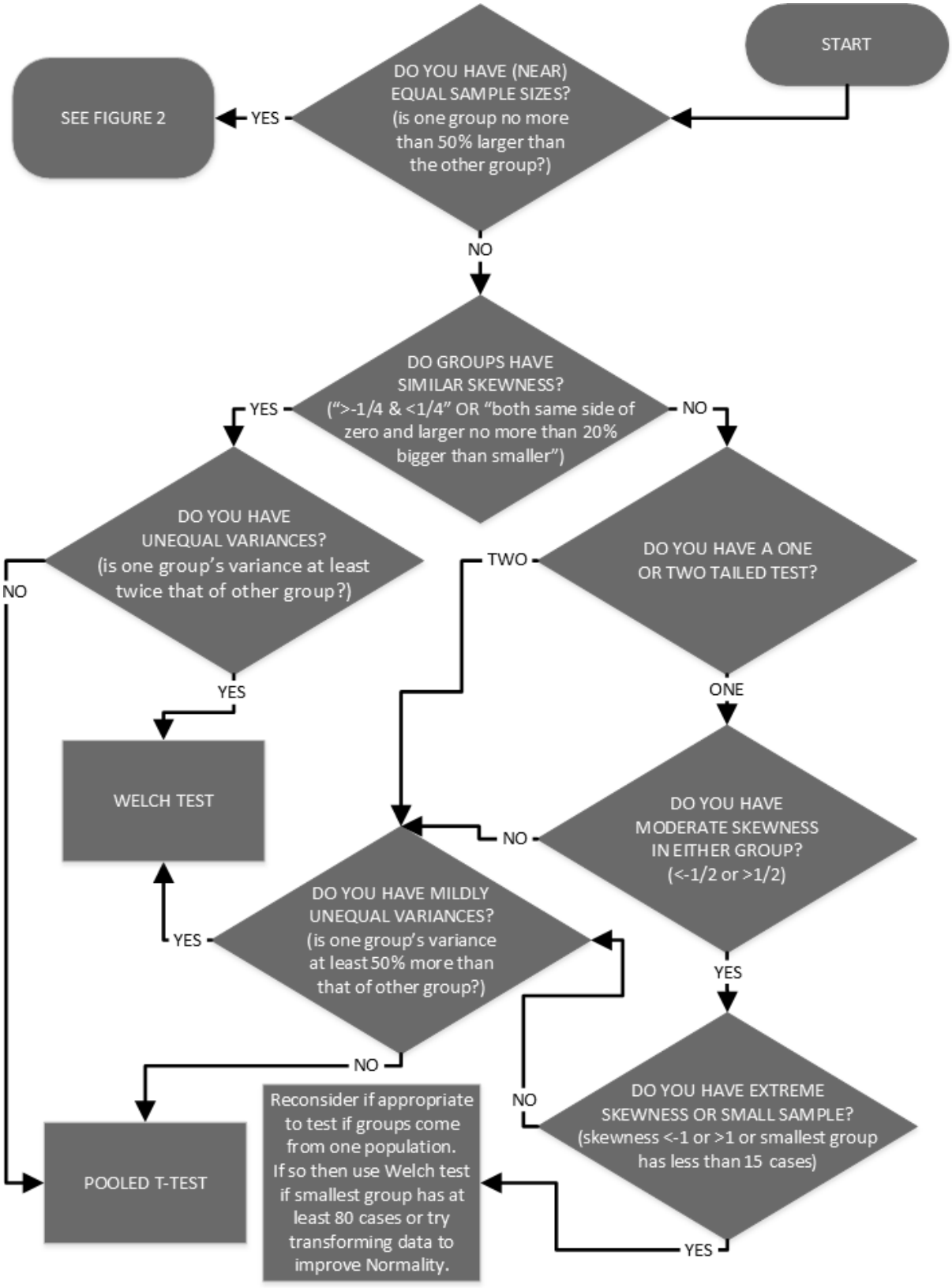

Figure 3: Flow Chart for Two-Sample Test (part 2)

It should be noted that the only tests to which the flow charts in Figure 2 and Figure 3 lead are the pooled t-test and Welch test (otherwise known as the separate variances t-test) and

there is no non-parametic test included. This is because only in situations so extreme that the entire analysis should be reconsidered are neither the pooled t-test nor Welch test appropriate.

Comparison of location for more than two samples

Issues to be considered for the comparison of location for several samples (more commonly referred to as parametric or non-parametric ANOVA) are similar to those for the two-sample test. However, with the absence of a parametric equivalent of Welch's test when variances cannot be considered to be the same, the choice of test is more limited.

Due partly to the historical roots of hypothesis testing in experimental design, the literature concerning how the parametric ANOVA performs in the presence of non-Normality is large. We summarize the conclusions that can be drawn from this literature below.

- Generally robust to departures from Normality (Pearson, 1931; Cochran, 1947; Gayen, 1950a; David and Johnson, 1951; Box, 1953).
- Lack of equality of variances is not a major problem unless the differences are major (Pearson, 1931, David and Johnson, 1951; Boneau, 1960) or differences in skewness also exist (Harwell, Rubinstein, Hayes, and Olds, 1992).
- Kurtosis more important than skewness but the latter can counteract the effects of the former (Gayan, 1950a). Only extreme kurtosis has a major effect (Boneau, 1960). When samples are of equal size, effect of kurtosis cancels out (Pearson, 1931). Kruskal-Wallis can be better than ANOVA when high kurtosis exists, particularly when sample sizes are large (Khan and Rayner, 2003).
- Similar skewness does not cause difficulties but different skewness can cause problems (Harwell et al, 1992). ANOVA is troubled by extreme skewness (Cochran, 1947).
- Equal sample sizes makes ANOVA more robust (Pearson, 1931; Boneau, 1960; Harwell et al, 1992) but not to extreme departures from Normality (Tan, 1982).
- With equality of variances and sample size of 32 or more in equal sized groups, ANOVA is insensitive to moderate departures from Normality (Donaldson, 1968).
- With large sample sizes, ANOVA is only sensitive to extreme skewness or kurtosis (Gayen, 1950a; Tiku, 1964).

We have translated these findings into a flow chart (Figure 4) to be used as practical guidance for researchers. As with previous Figures, decisions are made at various stages based on how equal the sample sizes are, the magnitude of skewness and kurtosis and how equal the variances are. Rationale for the cut-offs used follow below; some of which mirror decisions made for the two-sample tests above. We note that the Kruskal-Wallis non-parametric test assumes equality of variation; if this is not the case then a series of two-sample tests should be undertaken, adjusting for the inequality of variation and also for multiple testing.

Considering whether variances are similar or not, we again refer to Box (1954) and consider variances to be different if one is twice the size of that for another group or more.

When it comes to sample sizes being almost the same or different, Pearson and Adyanthāya (1929) and Box (1954) are the sources we again use. Once more we take a conservative approach and consider them to be different if one of the group's sample size is more than 50% larger than that of another. The work of Donaldson (1968) indicates that if each group has at least 16 cases, the overall dataset can be considered large enough to overcome moderate skewness or kurtosis. Gayen (1950a) and Tiku (1964) indicate that only if much larger sample sizes exist (i.e. at least 60 cases per group) can extreme skewness or kurtosis be ignored. In these circumstances, if the skewness and kurtosis is similar across

groups it would be more appropriate to use a non-parametric approach as the large sample size will give the test considerable power.

To decide whether skewness and kurtosis values are moderate or extreme, we refer to Pearson (1931). Geary (1947). Box (1953). Lindquist (1953). As a result, we again define extreme skewness as <-1 or >1 and moderate skewness as <-1/2 or >1/2 (without reaching the levels for extreme). For kurtosis we define extreme as a value <-3 or >3 and moderate as <-1/2 and >1/2. When it comes to deciding whether or not the skewness/kurtosis of the two groups is sufficiently similar, we again look at Geary (1947) and Lindquist (1953), extending their guidance to kurtosis. We thus consider an increase of less than 20% (equivalently a decrease of 16.667%) to represent similar skewness/kurtosis values (subject to them both being positive or negative). We also consider that if all groups have skewness/kurtosis less than ±1/4 then they are similar.

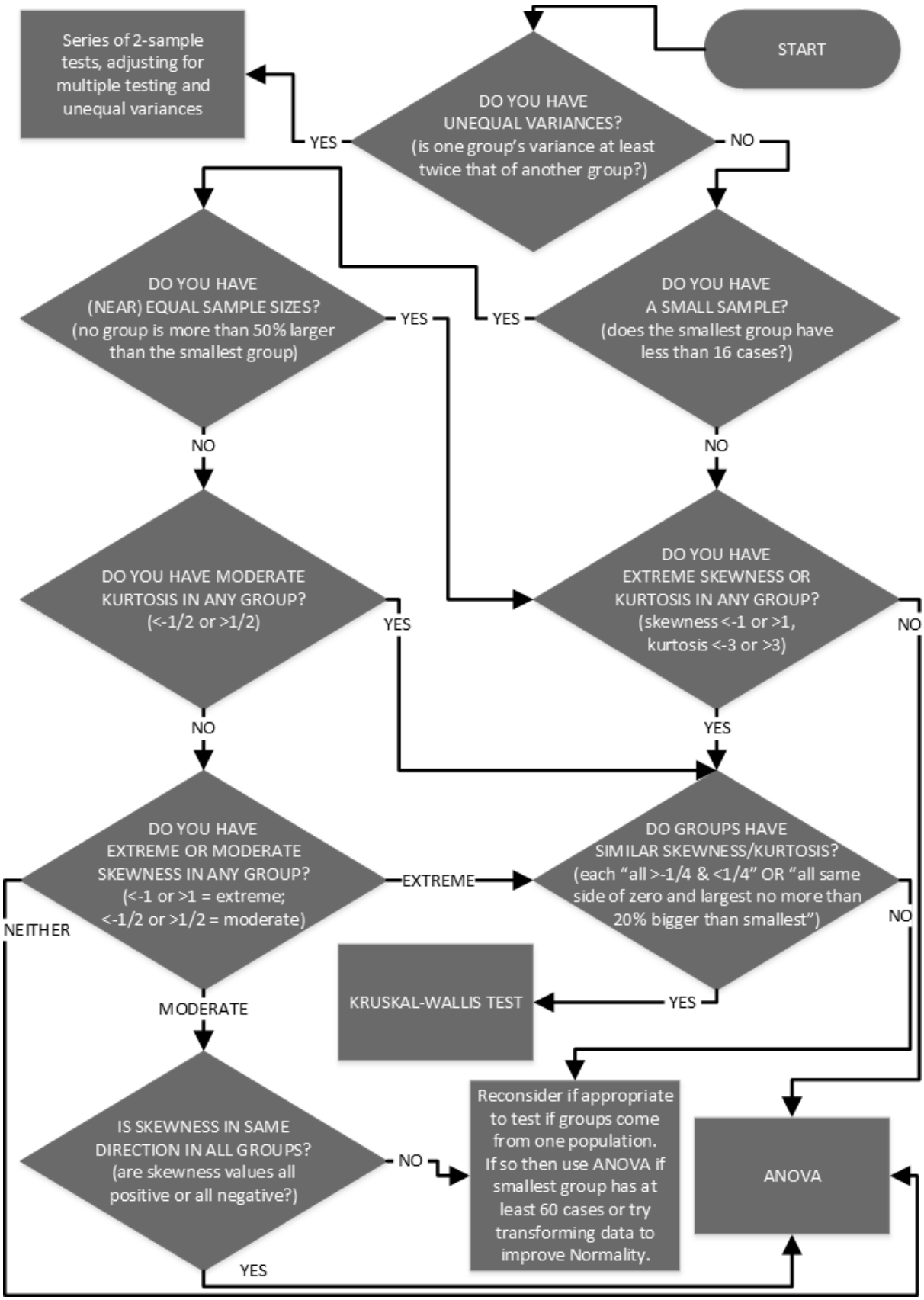

Figure 4: Flow Chart for Test of More than Two Samples

Discussion

We have demonstrated in this paper that advice regarding the issue of Normality available to researchers to date has been unsatisfactory. Either it is inappropriate (e.g. "conduct a formal hypothesis test for Normality") or vague (e.g. "look at a histogram"). Additionally we have shown that researchers, tasked with making appropriate choices when analysing their data, are concerned about the assumptions that are being made about the nature of the data. The issue of when data are "Normal enough" to use parametric tests has been addressed by renowned statisticians and there is a plethora of information in the literature, including summaries (e.g. Glass et al, 1972). However, until now, this had not been synthesized into an accessible format for researchers. In this paper we have sought to remedy this problem and produced flow charts that can help researchers decide on appropriate tests to perform.

We fully acknowledge that the act of translating the findings in the literature into flow charts is imperfect. In fact, it is inevitable that attempts to mould the advice of renowned statisticians such as Geary, with all its intricacies into a readily digestible format will always yield imperfections. However, we would argue that this is not sufficient reason for the task to be abandoned. Of course, p-values resulting from tests conducted as a result of following our flow charts should not be considered exact probability statements about the population of interest. Indeed, no researcher should believe the result of any hypothesis test to be so. The objective of the hypothesis test is to assess the evidence provided by the data and draw a conclusion as to whether it provides sufficient evidence to reject the null hypothesis. It is therefore important that the evidence is assessed using an appropriate test; the flow charts given in this paper help in this regard. The precise size of the p-value resulting from the test should be a secondary consideration after the decision has been made to accept or reject the null hypothesis and should perhaps only be published as an act of transparency rather than because of any substantive meaning placed upon it. Taking these issues into account, we do not believe that any imperfections caused by the translation of the advice from the literature into the flow charts will materially affect the outcome of the hypothesis test except in marginal cases where the outcome is more likely to be influenced by the effect of random sampling than the choice of test. Rather, the fact that appropriate decisions are being made as to the suitability of tests will lead to more correct conclusions being drawn.

It is a matter of debate as to whether hypothesis testing in the way it is traditionally conducted and used in research is the most appropriate for answering a research question. However, it is a fact that hypothesis testing is routinely used by research communities and will continue to be so for the foreseeable future. We would therefore argue that the statistical community has a duty to help researchers use these techniques through the provision of accessible guidance, such as that described in this paper.

References

Abu-Bader, S. H. (2010). *Advanced and multivariate statistical methods for social science research with a complete SPSS guide*, Chicago, Ill: Lyceum Books.

Barrett, J.P. & Goldsmith, L. (1976). When is n sufficiently large?, *The American Statistician*, 30, 2, 67-70.

Bartlett, M. (1935). The effect of non-normality on the t distribution, *Mathematical Proceedings of the Cambridge Philosophical Society*, 31, 02, 223-231.

Blair, R. C. and Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under various population shapes, *Psychological Bulletin*, 97, 1, 119-128.

Boneau, C. A. (1960) The effects of violations of assumptions underlying the t test, *Psychological Bulletin*, 57, 1, 49-64.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, 40, 3/4, 318-335.
Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. effect of inequality of variance in the one-way classification, *The Annals of Mathematical Statistics*, 25, 2, 290-302.
Cochran, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied, *Biometrics*, 3, 1, 22-38.
David, F. N. and Johnson, N. (1951). The effect of non-normality on the power function of the F-test in the analysis of variance, *Biometrika*, 38, 1/2, 43-57.
de Vaus, D. (2002). *Analyzing social science data: 50 key problems in data analysis*, London, UK: Sage.
Donaldson, T. S. (1968). Robustness of the F-test to errors of both kinds and the correlation between the numerator and denominator of the F-ratio, *Journal of the American Statistical Association*, 63, 322, 660-676.
Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). London, UK: Sage.
Fisher, R. A. (1925). *Statistical methods for research workers*, Edinburgh, UK: Oliver and Boyd.
Fisher, R. A. (1935). *The design of experiments*, Edinburgh, UK: Oliver and Boyd.
Fowler, J., Jarvis, P. and Chevannes, M. (2002). *Practical statistics for nursing and health care*, Hoboken, NJ: Wiley.
Gayen, A. K. (1950a). The distribution of the variance ratio in random samples of any size drawn from non-normal universes, *Biometrika*, 37, 3/4, 236-255.
Gayen, A. K. (1950b). Significance of difference between the means of two non-normal samples, *Biometrika*, 37, 3/4, 399-408.
Geary, R. C. (1947). Testing for normality, *Biometrika*, 37, 4/5, 209-242.
Glass, G. V., Peckham, P. D. and Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance, *Review of Educational Research*, 42, 3, 237-288.
Grove, S. K., Burns, N. & Gray, J.R. (2013). *The practice of nursing research: appraisal, synthesis, and generation of evidence* (7th ed.). St. Louis, Miss: Elsevier Saunders.
Harwell, M. R., Rubinstein, E. N., Hayes, W. S. and Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one-and two-factor fixed effects ANOVA cases, *Journal of Educational and Behavioral Statistics*, 17, 4, 315-339.
Howell, D. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury/Thomson Learning.
Khan, A., and Rayner, G. D. (2003). Robustness to non-normality of common tests for the many-sample location problem, *Journal of Applied Mathematics and Decision Sciences*, 7, 4, 187-206.
Lantz, B. (2013). The impact of sample non-normality on ANOVA and alternative methods, *The British Journal of Mathematical and Statistical Psychology*, 66, 2, 224.
Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education*, Boston, MA: Houghton Mifflin.
Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study, *Journal of Educational Measurement*, 7, 4, 263-269.
Markowski, C. A. and Markowski, E. P. (1990). Conditions for the effectiveness of a preliminary test of variance, *The American Statistician*, 44, 4, 322-326.
Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation, *Biometrika*, 23, 1/2, 114-133.

Pearson, E. S. and Adyanthāya, N. (1929). The distribution of frequency constants in small samples from non-normal symmetrical and skew populations, *Biometrika*, 21, 1/4, 259-286.

Ratcliffe, J.F. (1968). The effect on the *t* distribution of non-normality in the sampled population, *Journal of the Royal Statistical Society, Series C*, 17, 1, 42-48.

Salkind, N. (2014). *Statistics for people who (think they) hate statistics* (5th ed.). New York, NY: Sage.

Swift, L. and Piff, S. (2010). *Quantitative methods: For business, management and finance*, Basingstoke, UK: Palgrave Macmillan.

Tan, W. Y. (1982). Sampling distributions and robustness of t-ratio, F-ratio and variance-ratio in 2 samples and ANOVA models with respect to departure from normality, *Communications in Statistics Part A - Theory and Methods*, 11, 22, 2485-2511.

Tukey, J. W. (1948). Some elementary problems of importance to small sample practice, *Human Biology*, 20, 4, 205-214.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). San Diego, CA: Elsevier.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions, *The Journal of Experimental Education*, 67, 1, 55-68.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances, *British Journal of Mathematical and Statistical Psychology*, 57, 1, 173-181.

Zimmerman, D. W. (2011). A simple and effective decision rule for choosing a significance test to protect against non-normality, *British Journal of Mathematical and Statistical Psychology*, 64, 388-409.