That assumption of normality

Paul White,

Applied Statistics Group,

Faculty of Environment and Technology,

Univeristy of the West of England, Brsitol,

Bristol BS16 1QY, UK

paul.white@uwe.ac.uk

Paul Redford,

Department of Health and Social Sciences

Faculty of Health and Applied Sciences,

University of the West of England, Brsitol,

Bristol BS16 1QY, UK

paul2.redford@uwe.ac.uk

James Macdonald,

Department of Health and Social Sciences

Faculty of Health and Applied Sciences,

University of the West of England, Brsitol,

Bristol BS16 1QY, UK

james.macdonald@uwe.ac.uk

***Abstract*—** Many statistical tests are based around an assumption of "normality". The reasoning for this choice of distribution, whether to test, what to test, and how to test for normality is covered along with practical recommendations.

Keywords— Normal distribution, parametric tests, QQ plots, Central Limit Theorem, Preliminary tests, tests of normality

I. INTRODUCTION

Statistical tests are often labelled as "parametric tests", or "non-parametric tests" or "distribution free tests".

Parametric tests make an assumption about the functional form of an underlying distribution. This assumption allows the development of mathematical theory to draw inferences about a parameter of that distribution. The distribution could be, for instance, the normal distribution or the exponential distribution or one of many countless distributions. The parameters are constants that appear in the probability density function. However, in a wider sense other quantities or distributional properties such as the mean or median or variance could be regarded as a parameter of the distribution.

Distribution free tests are tests that relate to a parameter of a distribution (such as the mean or median) but the statistical test would be derived without specifying the underlying distribution. Randomisation tests and some bootstrapping tests would fall into this category.

Ranked based nonparametric tests are concerned with testing equality of distributions (rather than specific parameters) without specifying the functional form of the distributions.

The most commonly used parametric tests are statistical tests which are based on an assumption using the normal distribution, and example tests would include the independent samples t-test or the paired samples test, or the one-way between-subjects ANOVA.

It is worthwhile to note that a normal distribution is a theoretical distribution and perfect normality will not be obtained in any dataset. However, sample distributions of data might have characteristics which indicate that the underlying distribution might *approximate* a normal distribution. In these latter situations of *approximate* normality, the parametric tests that assume normality might work perfectly well. In fact, the normal assumed parametric tests might work reasonably well if sample data is clearly non-normal providing sample sizes are sufficiently large.

In this brief note, we will take a step back, and consider what is meant by a normal distribution [Section 2] and in Section 3 we will ask the question why an assumption of normality was made in the first place (i.e. why this distribution was picked on to develop theory and not one of the countless other distributions). A big part of all of this, is the word "assumption". An assumption is something, which is tentatively advanced, as opposed to a presumption, which is taken as a ground truth. We will therefore delve into whether the assumption should be examined, how might do this, and the consequences of doing so. Let's start with what is a normal distribution.

II. NORMAL DISTRIBUTIONS

A normal distribution is a **uni-modal symmetric distribution** for a **continuous random variable**, which, in a certain qualitative sense, may be described as having a bell-shaped appearance. Of course, bells come in a variety of shapes, so this might not be a good analogy. Normal distributions are also known as Gaussian distributions after Carl Friedrich Gauss who first described the distribution in 1809. Figure 1 gives a graphical illustration of the functional form of some example normal distributions.

There are two parameters which control the normal distribution; one parameter is the mean (which locates the central position of the distribution) and the other parameter is the standard deviation (which depicts the amount of spread in the distribution). If we know the mean and the standard deviation of a normal distribution then we know everything about it. Of course, there is an infinite number of values for either the mean or the standard deviation and as such there are infinitely many different normal distributions. However, it follows that if we know the mean of the distribution and if we know its standard deviation then we have precisely identified which normal distribution is being considered.

Note that any normal distribution is symmetric around the mean value (mean = median = mode), but not all symmetric distributions are normal distributions. Note that greater and greater deviations in either direction from the mean become increasingly less likely, and the degree of spread in a normal curve is quantified by the standard deviation. The two points at plus and minus one standard deviation from the mean are the points of inflexion of the normal distribution (i.e. the change between the curve being convex and concave).

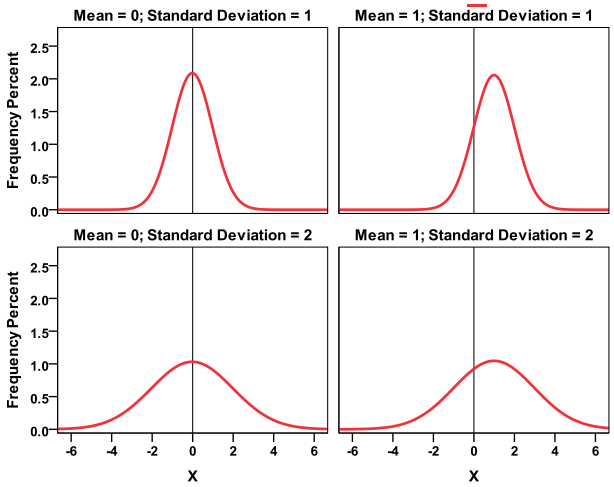

Figure 1 Probability density function for example normal distributions

Note that,

(a) the normal distribution is an example of a distribution for a theoretically continuous random variable [a continuous random variable is a random variable in which there are infinitely many values in any finite interval]

(b) the theoretical normal curve covers the entire real number line running from minus infinity to plus infinity

A moment's thought on these two points reveals that **a perfect normal distribution will not be encountered in any real practical context** arising in empirical research.

For instance, consider point (a). Suppose we are interested in the head circumference of neonates. Head circumference is a length, which, conceptually could be determined to any degree of precision by using better and better measuring equipment. In practice head circumference would be measured to the nearest millimetre. Accordingly, if recording neonatal head circumference then the data would be recorded to a finite number of decimal places and strictly speaking this data would be discrete (a discrete random variable is one in which there are a finite number of possible values in any finite interval). Of course, in practice, if the number of possible discrete outcomes was large and if the underlying measure is inherently continuous then we may argue that we are dealing with a continuous random variable and use statistical methods designed for continuous data without loss of accuracy.

Likewise, consider point (b). Again, suppose we consider neonatal head circumference. Clearly, we cannot have a negative head circumferences (but the normal distribution covers the negative number line) or very small positive head circumferences or very large head circumferences. In other words, in practice, there is a restricted range for neonatal head circumference. However, the normal distribution covers the entire number line and consequently neonatal head circumference could not have a *perfect* normal distribution.

Pedantic considerations of these aspects indicate that a perfect normal distribution will not be encountered in any real practical context arising in empirical research, and this is why Geary [1], as far back as 1947, suggested that the first page of every statistical textbook should contain the words "*Normality is a myth. There never was and will never be a normal distribution*".

However, this finite range restriction and the real word use of finite precision data does not invalidate the use of a normal *model* in a practical sense. A model in this sense is an attempt to describe a phenomenon of interest and is recognised to be an approximation (hopefully a good approximation) to reality. This idea is paraphrased by the statistician George Box who writes "*Essentially, all models are wrong but some are useful"* [2] and *" ... all models are wrong; the practical question is how wrong do they have to be to not be useful*" (ibid, p74).

III. WHY CONSIDER NORMALITY?

Many statistical tests are developed through postulating a statistical model composed of a systematic (aka deterministic or structural) component such as a trend or a difference and a random (aka stochastic or error) component to capture natural variation. Statistical scientists will make assumptions regarding the random component and then proceed to develop the best test for a given set of assumptions.

Many of the commonly used "parametric" statistical tests have been developed assuming the random component is normally distributed. Examples of these tests include t-tests, ANOVA tests, linear regression models, MANOVA, and linear discriminant analysis. In any practical situation, the assumption of normality will not be perfectly satisfied. However, computer simulations show that these commonly used parametric tests are **robust** to minor departures from normality. That is to say, these parametric tests still work very well in practice providing the assumption of normality has not been grossly violated. Moreover, in general it is fair

to say that increasing reliance can be placed on the validity of statistical conclusions from these tests with increasing sample size. This however does not answer the question "*why would a statistical scientist assume normality in the first place*?". The answer to this lies in a theorem known as the Central Limit Theorem.

IV. THE CENTRAL LIMIT THEOREM

Here comes the technical bit! Imagine a process whereby a random sample of size *n* is taken from a distribution that has a finite mean μ and a finite standard deviation σ (let's call this the parent distribution). The mean of this sample, $\bar{X}_1$, could be recorded. Now consider repeating this process taking another random sample from the parent distribution, again with the same sample size *n*, and again with the mean of the second sample being recorded $\bar{X}_2$. Conceptually this process could be repeated indefinitely giving a series of means $\bar{X}_1, \bar{X}_2, \bar{X}_3, \bar{X}_4, \bar{X}_5$ …., each based on the same sample size *n*. We might ask the question: "*What distribution could approximate the distribution of the sample means?*" (This is the child distribution.) The answer to this question is that, irrespective of the functional form of the original parent distribution, we have the following results:

i. the expected value of the means $\bar{X}$ is μ where μ is the mean of the original parent distribution (this seems reasonable, i.e. the average of averages is the average)
ii. the standard deviation of the means is $\sigma/\sqrt{n}$ where σ is the standard deviation of the original parent distribution; this seems reasonable too, since $\sigma/\sqrt{n}$ will tend towards zero as *n* increases; i.e. averaging is a smoothing process and with very large samples we would expect a sample mean to closely reflect the true theoretical mean and hence with large samples, the sample means would closely cluster around the true mean much more closely than the clustering of individual observations or means based on small samples
iii. the distribution of the means can be *approximated* by a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$

In point iii) the quality of the approximation depends on both the functional form of the original parent distribution and on the sample size. If the parent distribution is a normal distribution then the child distribution is also normal. Statistical simulations show that if the parent distribution is quite heavily skewed then sample sizes of $n > 60$ may be needed for means to have an approximate normal distribution; if the original parent distribution is moderately skewed then sample sizes of $n > 30$ might be needed for the means to have an approximate normal distribution; if the original parent distribution is symmetric then the approximation may still be deemed a good approximation with sample sizes smaller than $n = 30$. Of course, in practice, a researcher will only have one data set and therefore one mean. However, by virtue of the Central Limit Theorem, this mean can be considered to be a sample from a distribution which approximates the normal distribution and the quality of the approximation can be gauged by the above rules of thumb.

This is all well and good, but more importantly it is the *consequence* of the Central Limit Theorem (i.e. averages have a distribution which can be modelled using a normal distribution) which motivates theoretical statisticians to make a normality assumption in deriving what are now commonly used parametric statistical tests. For instance, consider neonatal head circumference. Neonatal head circumference for an individual is likely to be influenced by *many naturally varying factors* e.g. genetic or hereditary factors, nutritional factors, environmental factors and so on, including factors we might not know about. If these factors act in an independent additive manner, then this will induce variation across a population producing an *averaging effect* over individuals and hence by the Central Limit Theorem we would not be overly surprised if the resulting distribution could be approximated by a normal distribution without loss of too much accuracy. In other words, in a relatively homogeneous population, an outcome measure, which is affected by a large number of unrelated equipotent factors, will produce a distribution with some central target value (the mean) with extreme values consistently occurring less frequently. This might take some time to digest! The point is, under certain conditions there is prior reasoning to expect some outcome measures to be normally distributed and it is this reasoning that motivated the development of so many tests predicated on an assumption of normality. Examples of this would include height of boys aged 8 to 9, or weights of packets of crisps.

V. ASSESSING NORMALITY

There are three main approaches for assessing normality. In this note these approaches will be referred to as "mental imagery", "graphical and descriptive" and "formal inferential"

Mental Imagery

The first thing to do when assessing data for normality is to simply ask the question "how would I imagine the data to look?". This should be done prior to data collection. Some

simple reasoning about the form of the data might lead to an outright rejection of using a normal probability for that data. Some examples will make this clear.

Example 1

Suppose we are interested in the obstetrical history of women aged 16 to 21 and wish to record parity (i.e. number of pregnancies beyond 20 weeks gestation). Parity of each woman would be recorded; for each woman we would record whether they are nulliparous (parity = 0), whether they have been pregnant once beyond 20 weeks (parity = 1), whether they have been pregnant twice beyond 20 weeks (parity = 2), and so on, (i.e. for each woman there would be a count of either 0, 1, 2, 3, …). Now visualise the distribution of data that is likely to be collected for this population of women aged 16 to 21. Would you expect this data to be normally distributed? Of course, you would not. In all likelihood, the most frequent parity recorded for this population would be parity = 0 (nulliparous women), followed by parity = 1 (one pregnancy), followed by parity = 2 (two pregnancies). At the outset we would argue that we have a highly discrete distribution (taking numbers 0, 1, 2, 3, 4, 5 and fractional numbers e.g. 1.53 would not be possible), with a very restricted domain (e.g. the count cannot be negative and high numbers would be impossible), and that the distribution would be skewed to the right (aka positively skewed). Therefore, the distribution is discrete arising from counting whereas the normal distribution is for an inherently continuous variable usually arising from measurement. The domain of the distribution is over a very restricted range whereas the normal distribution is unrestricted. The distribution is positively skewed but the normal distribution is symmetric. These reasons would suggest that parity is not normally distributed. [As an aside, lack of normality does not mean that parametric tests such as t-tests cannot be used as other considerations; they still might be appropriate as discussed later.]

Example 2

Suppose we worked in a factory which produces nails with a target length of 50mm. Length is an inherently continuous measurement. We do not expect all of the nails to be exactly 50mm (or even any one of them to be exactly 50mm) instead we would expect some natural variation. We could anticipate a mean value of about 50mm with some lengths above 50mm, some lengths beneath 50mm, and unusually large deviations away from 50mm being less frequent. If you visualise the histogram of the above lengths you will obtain something resembling the classic church-bell shaped curve and in this instance, the assumption of normality might not seem too unreasonable to make. In this case, we would not be too surprised if the data turned out to be *approximately* normally distributed. This example also suggests we could use graphical techniques to help assess normality.

Graphical Techniques ("Chi-by-Eye")

A popular way to assess normality is to "eyeball" the data. John Tukey is a strong advocate for always producing graphical displays writing "*there is no excuse for failing to plot and look*" and specifically argues that graphical methods are a "*useful starting point for assessing the normality of data*" (see [3]). One commonly used graphical approach is to create a histogram of the sample data and to use the histogram to make a subjective appraisal as to whether normality seems reasonable. Figure 2 is an example histogram for some computer-generated data sampled from a normal distribution.

If data has been sampled from a normal distribution, then we might expect the shape of the sample data to be symmetric with the highest frequency in the centre and lower frequencies heading towards the extremes of the plot.

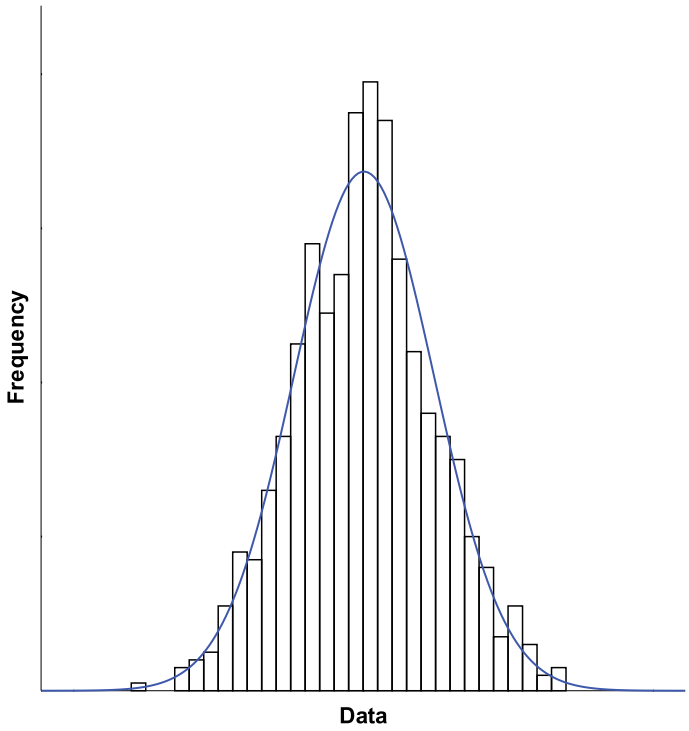

Figure 2: Histogram of normal data including the normal curve

However, there is a problem with histograms. Firstly, it is commonly recognised that the shape displayed in a histogram can be highly dependent on the histogram class width and the location of histogram boundaries. Changing class width or changing the class boundaries can greatly alter the shape of the histogram particularly when dealing with samples of size $n < 100$. Secondly, there is some doubt about the validity of subjective human assessments of histograms for judging

normality. For instance, suppose you had the time and inclination to write a computer program to generate say 1000 data sets each of size n = 50 each taken from a theoretical normal distribution and for each of these data sets you create a histogram (i.e. 1000 histograms). Inspection of these 1000 histograms would then give you some indication of the natural variability in histogram shapes that could be obtained when dealing with samples of size n = 50. By way of example, Figure 3a gives four sample histograms each based on n = 50 with all data sampled from a normal distribution. Do the histograms in these panels look as if they represent data sampled from a normal distribution? Would other people make the same judgement? The same data is also given in Figure 3b, this time with a different number of histogram bins. Do these histograms suggest the data has been sampled from a normal distribution?

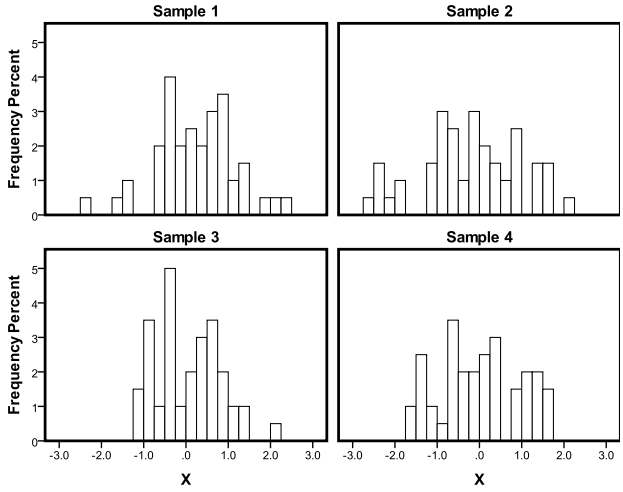

Figure 3a Each histogram is based on n = 50 sampled from the standard normal distribution.

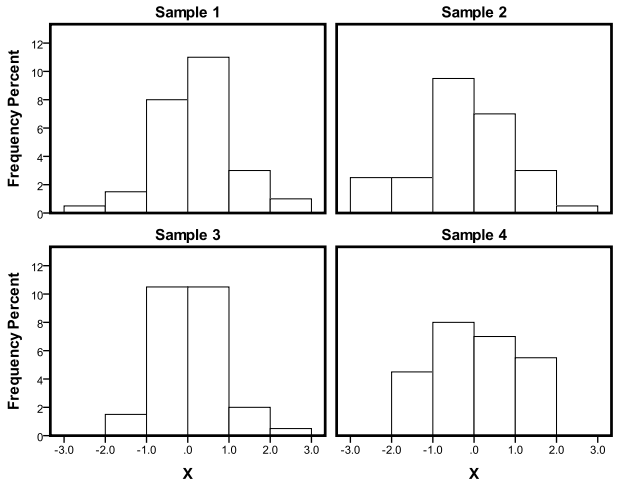

Figure 3b Each histogram is based on n = 50 sampled from the standard normal distribution.

In general histograms could be used to help form an opinion on normality but the visual effect of histograms are themselves dependent on chosen bin widths and we might not be trained to known what we are looking for.

In Figure 3b, the number of histogram bars is small (b = 5). What would be the minimum number of bars needed to capture detail? Perhaps 10, or 20? For arguments sake, let's say 10. There is no one fixed rule for the number of histogram bars (bins) in a histogram. One rule of thumb is the number bars, b, is $b = \sqrt{n}$ (the square root rule). Hence, under this rule we would need a sample size, n, of at least 100 justify using 10 histogram bars. Alternatively, the rule developed at Rice University [4] is to have $b = 2 \times \sqrt[3]{n}$ (i.e. 2 times the cube root of n). For b = 10, this implies a minimum sample size of n = 125. Likewise, using the rule given by Sturges, $b = log_2 n + 1$ which implies for 10 histogram bars a minimum sample size would be n = 512. The rule by Rice is probably better than the rule given by Sturges, but either way, these rules indicate that sample sizes of in excess of 100 are needed to justifiably have at least ten histogram bars, and that might be the minimum number of bars needed to see the detail in a histogram.

For these reasons a number of practitioners would inspect a box-and-whiskers plot (aka a "box plot") to help form an opinion on normality rather than using a histogram. Broadly speaking, a box plot is a graphical representation of the five-figure summary (minimum, lower quartile, median, upper quartile, maximum) of a sample distribution. The box-and-whiskers plot greatly assists in determining whether a sample is skewed and in screening for the presence of potential outliers. Detailed information on the creation and interpretation of box-and-whisker plots is given by Tukey [3] and will not be covered here.

A box plot created from a normal distribution should have equal proportions around the median. For a distribution that is positively skewed the box plot will show the median and both of its quartiles closer to the lower bound of the graph, leaving a large line (whisker) to the maximum value from the data. Negatively skewed data would show the opposite effect with the majority of points being in the upper section of the plot boundaries. It is expected that some outliers will occur which are shown by points either ends of the whisker lines. Figure 4a gives some sample box-plots for the normal distribution, a positively skewed distribution, a negatively skewed distribution and a distribution which has a very large central peak with very few observations in the tail of the distribution (i.e. a "peaked" distribution with a high degree of kurtosis).

Box-and-whisker plots are good visual devices for assessing *symmetry* in a distribution and this is a property of

the normal distribution (but not the only property). These plots also allow outliers to be quickly spotted. A major drawback of the box-and-whisker plot is that it does not readily convey information on sample size. An alternative graphical display to overcome the limitations of histograms (and to a lesser extent the limitations of box-plots) is the normal probability plot. A normal probability plot comes in two flavours: - either the Q-Q plot (quantile-quantile plot) or the P-P plot (percentile-percentile plot).

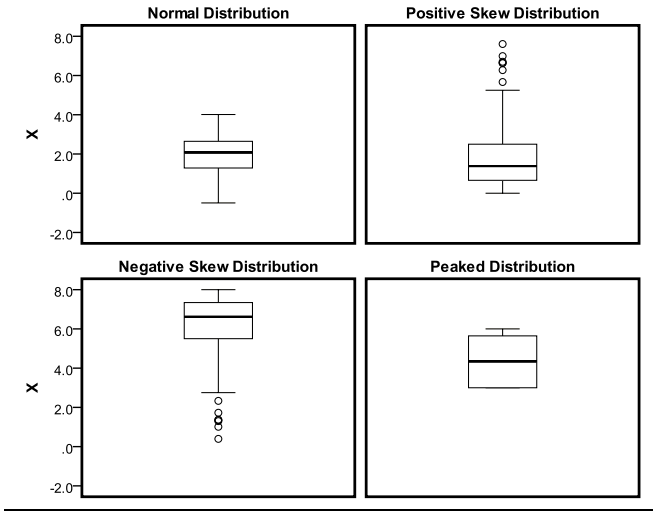

Figure 4a: Box Plot for normal data (upper left quadrant), positively skewed data (upper right quadrant), negatively skewed data (lower left quadrant) and a peaked distribution (lower right quadrant).

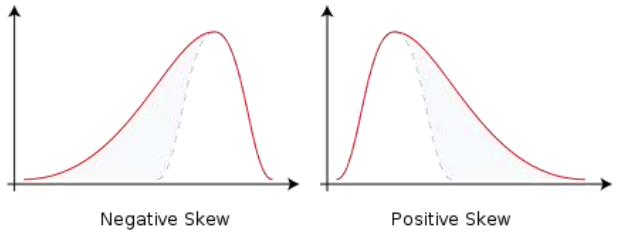

Figure 4b Schematic representation of distributions displaying a noticeable degree of skew.

Gaussian Q-Q and P-P Plots

The most common graphical tool to assess the normality of the data is a Quantile-Quantile (Q-Q) plot [5]. In a Q-Q plot, the quantile values of a theoretical distribution are plotted against the quantile values of the observed sample distribution (x axis). In a normal Q-Q plot the quantiles of the theoretical normal distribution are used. Thereafter the aim is to make a judgement as to whether the two quantiles are produced from the same distribution; if this was the case then the plotted points would create a straight diagonal line. Any systematic deviations from a straight line, other than natural random fluctuations, suggest that the distributions cannot be considered to be the same.

Closely related to the normal Q-Q plot is the normal percentile-percentile plot (P-P plot) which is a plot of the theoretical percentiles of a normal distribution (y-axis) against the observed sample percentiles (x-axis). If the sample data has been sampled from a normal distribution then, like the Q-Q plot, it is expected that the plotted points will fall along a straight line. If the data has been sampled from a non-normal distribution then systematic deviations from this line are expected (e.g. banana shaped plots for skewed distributions or S-shaped plots from distributions with tails which differ from the tails of the normal distribution. Figure 5 gives example P-P plots for the data previously displayed in Figure 4a.

Normal Q-Q and Normal P-P plots are preferred to histograms and box-plots for helping to make a subjective assessment of normality. Histograms suffer from an element of arbitrariness in choice of bins, possibly being sensitive in visual appearance to bin choice, and from not having a reference capturing what can be expected within the confines of natural sampling variation (although superimposing a best fitting normal curve on the histogram would helpfully assist interpretation). Similarly, box-plots are excellent for judging symmetry but symmetry is not the only feature of a normal distribution. In contrast the Normal Q-Q plot and the Normal P-P plot are specifically designed to visually assess normality and incorporate a theoretical normal distribution in their creation. However, it is conceded that both Normal Q-Q plots and Normal P-P plots are open to subjective interpretation. For these reasons, some may want to statistically test for normality using an inferential test.

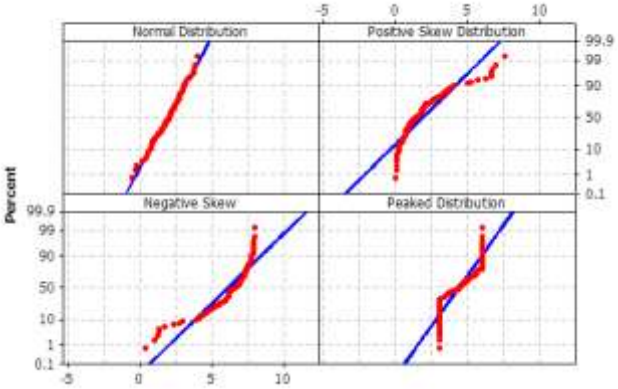

Figure 5 Normal (Gaussian) P-P plots for normal data (upper left quadrant), positively skewed data (upper right quadrant), negatively skewed data (lower left quadrant) and a peaked distribution (lower right quadrant).

It is worth noting that lack of normality is often shown in the tails of a distribution. The Normal -Q-Q plot would tend to pick this up. In contrast the Normal PP plot is constrained between 0% and 100% and is, therefore, less likely to show

deviations from the tails of a normal distribution. This precious distinction is worth bearing in mind. [Cynically, it has been suggested to use Normal Q-Q plots to show non-normality and Normal P-P plots to show normality!]

Tests of normality

There are countless tests of normality. Example tests include the Kolmogrov-Smirnov test, or its modification known as Lillefor's test, or the D'Agostino test, or the Jarque-Bera test, or the Cramer von Mises test, or the Shapiro-Wilk test, or the Epps and Pulley test. The list goes on. There is no one single "best test" for testing normality and there never will be. The monograph [6] compares and contrasts the properties of 40 tests of normality but even this monograph does not provide comprehensive coverage and it omits a number of normality tests that are well known to the statistics community.

In testing for normality, the statistical hypotheses are of the form:

S_0 The data are an independent identically distributed (*iid*) random sample from a normal distribution

S_1 The data are not *iid* normally distributed

or

H_0 Underlying distribution is a normal distribution with some unknown mean μ and some unknown variance σ^2

H_1 The underlying distribution is not a single normal distribution.

In practice the main use of tests of normality is to investigate whether assumptions underpinning the so called "parametric" tests are justifiable. Often there is a strong desire by the research community to use standard parametric tests and in these cases a researcher would be looking for a confirmation of the appropriate normality assumption. In these situations, the researcher would not want to reject the null hypotheses as stated above. However, if we take the view that a perfect normal distribution will not be encountered in any real practical context then it follows that H_0 must be false. Indeed, if normality does not exist in practice and if we take a sufficiently large sample then statistical tests of normality will lead to the rejection of H_0. On the other hand, a failure to reject H_0 would be a Type II error!

The above problem is compounded further by the general desire to have good powerful statistical tests. Accordingly, statistical scientists have developed tests such as the Lin-Mudholkar test of normality which is very powerful for detecting lack of normality when the distribution is skewed, or the Shapiro-Wilk test which is very powerful when sample sizes are < 50, or the Jarque-Bera test which is powerful for detecting changes in skewness and/or kurtosis, and so on.

A question that we can consider is "*Do we really want to use a test of normality which is powerful*?" i.e. do we want to use a test which is very good at detecting lack of normality and therefore having a high chance of rejecting H_0? We might, we might not. From a theoretical perspective the parametric tests such as t-tests, regression, ANOVA, etc are the best tests available if data is normally distributed and in general these tests are robust to minor departures from normality. Accordingly, if assessing assumptions for normality then there is a line of reasoning to use a statistical test of normality which will pick up large departures from normality but be less sensitive to minor deviations from normality. This line of reasoning suggests using a valid test but one which is not overly powerful. One such test is the Kolmogorov-Smirnov test which can be used to statistically test for normality.

Given the robustness of the parametric t-tests and similar, it would be preferable to test a null hypothesis

H_0 Underlying distribution is *approximately* normal

However, this is not a point null hypothesis; the word "approximately" is too vague. In null hypothesis testing, *there is no formal statistical test of approximately normal.*

If there is a desire to test for normality then a common problem is a failure to understand what to precisely test. What precisely should be "normal"?

VI. WHAT SHOULD BE TESTED? THE IID ASSUMPTION

The commonly encountered parametric statistical techniques (e.g. independent samples t-test, one-way between subjects ANOVA, two-way between subjects ANOVA, linear regression etc) have a theoretical development based on assumptions that errors are independent identically distributed normal random abbreviated to "*iid* normal". For instance, in a linear regression model, the model under consideration would have the form

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

with the assumption that the error terms, ε_i, are independent identically distributed (*iid*) normal random variables with a mean of zero, i.e. $\varepsilon_i \sim N(0, \sigma^2)$. Note that we are NOT saying the data (x) is (*iid*) normal but we are making assumption that the *errors* are *iid* normal random variables with a mean of zero and some unknown standard deviation σ. In any practical situation, the errors will be unknown and they will be approximated by the sample residuals.

For another example, consider the one-way between-subjects ANOVA. The model usually used for the one-way ANOVA has the form

Testing that assumption of normality

$$Y_{i,j} = \mu + \tau_j + \varepsilon_{i,j}$$

where $Y_{i.j}$ is the outcome (dependent variable) for the *i*-th observation in the *j*-th group, where μ denotes some overall mean, τ_j denotes the effect of the *j*-th group, and ε_{ij} denotes a random error for the *i*-th observation in the *j*-th group. For development purposes, the error terms, ε_{ij}, are assumed to be *iid* normal (and not the "data"). For instance, suppose a computer package is used to generate 100,000 independent normal random deviates, sampling from a normal distribution with mean zero and a standard deviation 1. Let's call this data "data from group 1". A histogram of the group 1 data is given in Figure 6. Note that the data in Figure 6 is an example of *iid* normal data (i.e. each data point is independent of any other data point, and each data point has been sampled from the same normal distribution).

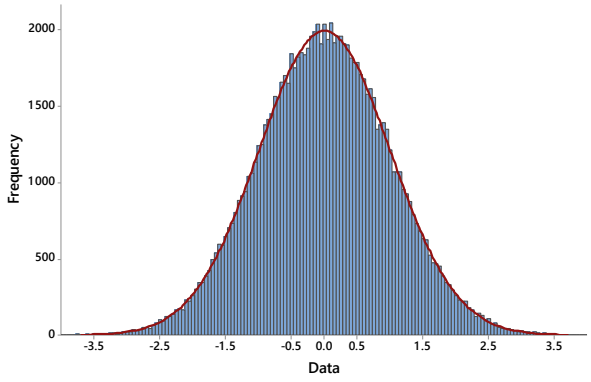

Figure 6 Histogram of n = 100,000 *iid* normal deviates

Let's suppose this computer exercise is repeated but this time taking a sample of n = 100,000 but from a normal distribution with a mean of 3 and standard deviation of 1. Again, this sample (data from group 2) would be *iid* normal. Further suppose this exercise is repeated but this time taking a sample of n = 100,000 but from a normal distribution with a mean of 6 and standard deviation of 1. Again, this sample (data from group 3) would be *iid* normal. And finally, suppose this exercise is repeated, again taking a sample of n = 100,000 but this time from a normal distribution with a mean of 12 and standard deviation of 1. Again, this sample (data from group 4) would be *iid* normal. Figure 7 shows the four sample histograms; in each group the data certainly look to be normally distributed. Figure 8 is a normal PP-plot for the four sets of data and this graphic too aligns with the notion that the data in each group is *iid* normal.

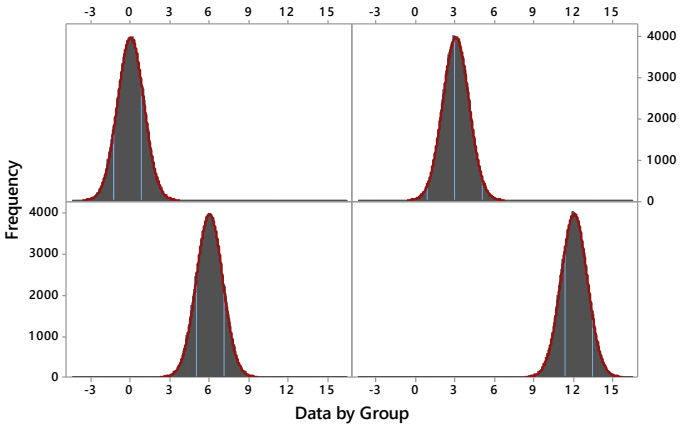

Figure 7 N = 100,000 normal deviates with constant variance but with differences in location

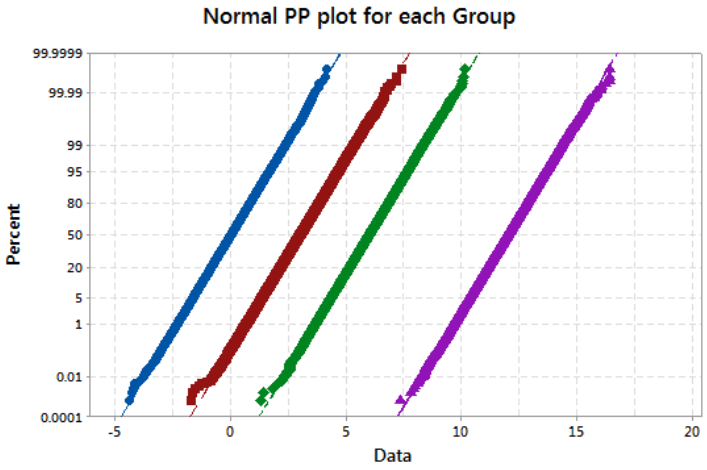

Figure 8 Normal P-P plot with N= 100,000 deviates per group

Now suppose the four sets of data are put together into one data set. The resulting data is shown in Figure 9a and Figure 9b.

A cursory inspection of Figure 9a or Figure 9b would indicate that the "data" is not *iid* normal (in fact the data is from a mixture of four different normal distributions). The normal probability plot of the combined sample (Figure 10) clearly shows non-normality.

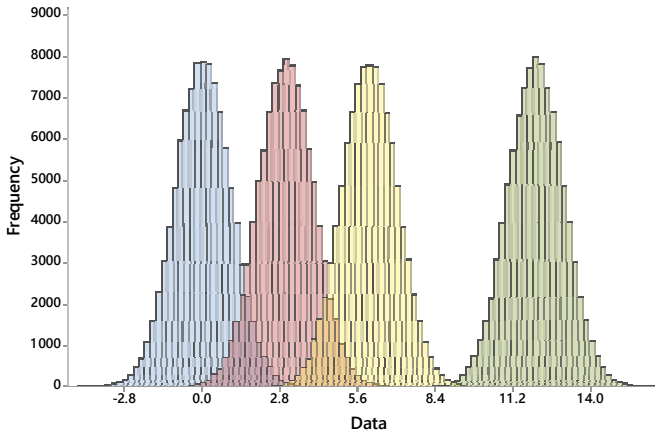

Figure 9a Histogram with N = 100,000 deviates per group

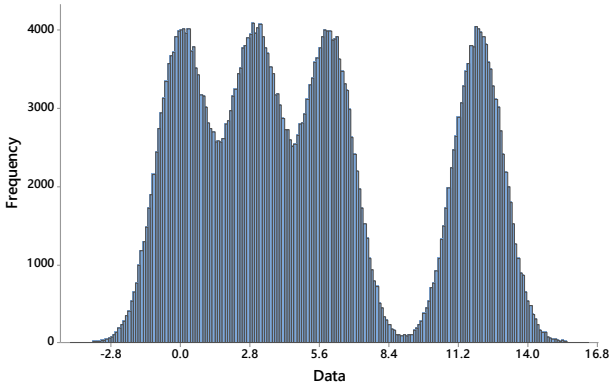

Figure 9b Histogram of N = 400,000 deviates (a mixture distribution)

All of this demonstrates that data combined over different groups may not be *iid* normal even if data within each group is *iid* normal. Consequently, if aiming to "test" or examine the assumptions underpinning an ANOVA (or a regression, or a t-test) then it is **not** the combined data that should be assessed for normality. Instead, it would be either (a) examine each group separately for *iid* normality or (b) examine the residuals for normality. Approach (a) involves multiple testing or multiple examining . Approach (b) is the preferred approach; after all that is the statistical assumption being made.

The point is, there is a need to be precise over what is assumed to be normally distributed. The parametric statistical tests have an underlying mathematical (systematic, deterministic, structural) component coupled with a statistical (stochastic, error, random) component. This later component is quantified by "residuals". The takeaway message, is, i*f an assessment of normality is being made then the assessment is done on the residuals and not the "data".*

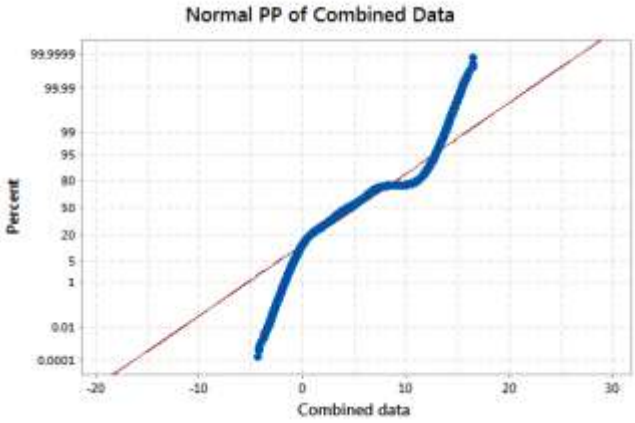

Figure 10. N= 400,000 deviates from a mixture distribution

VII. PRELIMINARY TESTING

In general, statistical tests performed to test assumptions and to inform the choice of the analytical technique for the main analysis, are known as preliminary tests. An example preliminary test, is when Levene's test is used to assess equality of variance between two independent scale samples, with a view to determining whether to use the independent samples t-test or the separate variances t-test (aka Welch's test). Another example of a preliminary test would be to test for normality in two independent samples to help decide as to whether to analyse the data assuming normality (e.g. independent samples t-test) or not (e.g. use the Mann Whitney Wilcoxon test). There is some debate over this practice.

Some analysts would use preliminary tests with a nominal alpha = 0.05 significance level. It is not clear whether this is a sensible choice. The logic of preliminary hypothesis testing of assumptions is different from the logic of drawing scientific conclusions in superiority contrasts. For instance some might argue, because of the robustness of some parametric tests to violation of assumptions, then they might only consider an underpinning assumption to not be tenable if they obtain a significant result at the alpha = 0.001 level (or lower). Or, alternatively, if using Levene's test to effectively choose between a default position of the independent samples t-test (i.e. equal variances assumed) or Welch's test then an alpha = 0.20 (or higher) might be used arguing no harm is done if Welch's test is used (and also arguing that an assumption of equal variances should not necessarily be the default). In summary, amongst those who assess assumptions using formal statistical tests there is no consensus on the nominal significance level to be used.

There is also an unintended statistical consequence of preliminary testing. Preliminary testing is a way of letting the data select the test that will be used in its analysis; this

process can adversely affect the statistical error rate. For instance, suppose we consider a situation where we have two independent groups with scale data. An analyst might employ a strategy of formally assessing the sample data for normality and

(a) if a statistically significant result is not obtained then use the independent samples t-test to compare the two samples or

(b) if a statistically significant result is obtained (i.e. a statistically significant departure from normality in the sample) then use the Mann Whitney test to compare the two samples.

Under this strategy, *the analyst is allowing the data and preliminary test to pick the statistical test*.

Suppose we use computer simulation to analyse this strategy. In our simulation, we will sample from normal distributions. In our simulation, we will make the null hypothesis of equal means to be true too. There are two situations to consider.

Situation A. …. In any one instance, we fail to reject the null hypothesis "the errors are *iid* normal" and in these cases we proceed to use the independent samples t-test to test for mean differences.

Situation B … We reject the null hypothesis "the errors are *iid* normal" and in these cases we proceed to use Mann Whitney test to test for distributional differences in location.

Unfortunately, in Situation A, the Type I error rate [false positive rate] for the resulting independent samples t-tests would be lower than the normal Type I error rate.

Unfortunately, in Situation B, the Type I error rate [false positive rate] for the resulting Mann Whitney tests would be much higher than the normal Type I error rate.

Fortunately, the over and under estimation in Situation A and Situation B tend to cancel one another out. However, it does remain an unintended consequence that error rates and power can be affected by preliminary testing. Some have suggested that if formal testing does cause these unintended consequences then perhaps formal testing should not be done and informal assessments, such as QQ plots be used instead. However, this does not overcome the problem; it merely replaces one set of formal tests with another set of informal tests. How can this perceived problem be resolved? A partial answer is to plan, plan and plan beforehand i.e. prior to any data collection, to really think through all of the relevant analyses and justify them; the more that can be pushed upstream to a formal statistical analysis plan the better. Of course this is not always possible but there are benefits to actively staying in the statistical analysis stage.

VIII. SUMMARY

The consequences of the Central Limit Theorem suggest that approximate normality is likely to occur in some facets of nature and society. Approximate normality is conceptually useful but it presents mathematically intractable challenges for developing theory. For this reason, the mathematical development of the commonly used parametric tests is based on an assumption of (precise) normality. It turns out that the derived statistical tests are not overly dependent on this precise assumption being satisfied. In any event, t-tests and ANOVA only require means to be normally distributed or approximately so. This latter requirement might be satisfied if the parent distribution is symmetric and sample sizes Are between 10 and 30 per subgroup; or sample sizes of 30 and above per subgroup for mild degrees of skewness; or sample sizes greater than 60 per subgroup for moderate levels of skewness.

Sometimes there is a need to assess normality. The assessment should be done on the residuals. The null hypothesis concerning normality is a precise null hypothesis and it is always wrong. For this reason there are calls to not formally test for normality. Certainly, the null hypothesis concerning normality would be rejected if sample sizes are sufficiently large. There is an argument that such formal tests might have some merit if sample sizes are small and the tests are simply being used to screen for large non-normality. The use of normal Q-Q may be a better way of appraising the extent of departure from normality. However, both formal and informal assessments of normality are forms of preliminary testing. Preliminary testing could be helpful in avoiding the use of an undesirable test but could also have an unintended consequence by failing to adequately control statistical error rates. The best remedy for this, is to carefully construct the study investigation and statistical analysis plan at the outset. This does not detract from exploring the data (and there may be a scientific duty to do a full forensic examination of the data, and a moral duty to your participants to fully examine the data set). However, the more pre-study preparation the better and this will minimize the problems associated with that assumption of normality.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of the Learning and Teaching Initiative in the Faculty of Health and Applied Sciences, University of the West of England, Bristol in supporting the wider Qualitative and Quantitative Methods teaching programme.

SERIES BIBLIOGRAPHY

White P, Redford PC, and Macdonald J (2019) A primer on validity and design terminology in comparative designs, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –4

White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the chi-squared test of association (2 by 2), *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –4

White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the independent samples t-test and the Welch test, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –6

White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the paired samples t-test, *Qualitative and Quantitative Research Methods Project,* University of the West of England, 1 –5

White P, Redford PC, and Macdonald J (2019) A primer on statistical hypotheses and statistical errors, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –4

White P, Redford PC, and Macdonald J (2019) A reverse look at p-values, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –5.

White P, Redford PC, and Macdonald J (2019) That assumption of normality, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –11

White P, Redford PC, and Macdonald J (2019) Cohen's *d* for two independent samples, *Qualitative and Quantitative Research Methods Project,* University of the West of England, 1 –4

REFERENCES

[1] Geary R C (1947) Testing for normality. *Biometrika*, 34, 209 – 242,

[2] Box G E P and Draper NR (1987) *Empirical Model-Building and Response Surfaces*, John Wiley and Sons

[3] Tukey J W (1977) Exploratory data analysis. Reading MA: Addison-Wesley

[4] Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University

[5] Razali N M and Wah, Y B (2011) Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lillefors and Anderson Darling tests, *Journal of statistical Modeling and Analytics*, Vol 2, No 1, 21 – 33.

[6] Thode HCJ. Testing for normality. New York' *Marcel Dekker*, Inc.; 2002. p. 1– 479.