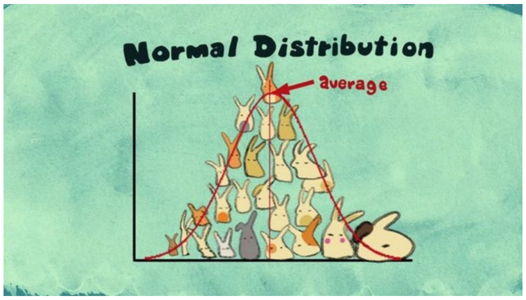
Order Statistics and Normal Probability Plots



J N S Matthews
Biostatistics Research Group,
Newcastle University

1 Some useful revision

In this article use will be made of the properties of Beta variables and a few other results. For convenience, a reminder of the key features are given here.

1.1 The Beta distribution

A Beta random variable is a random variable that is in the interval (0,1) and is defined by means of two positive parameters, α and β, and denoted by Beta(α, β). It has density

$$f(x \mid \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \qquad 0 < x < 1$$

where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$ is the Beta function. The mean of the distribution is $\alpha/(\alpha + \beta)$ and its variance is $\alpha\beta/[(\alpha + \beta)^2(\alpha + \beta + 1)]$.

1.2 Sample mean and variance

For a sample of independent, identically distributed Normal random variables, the sample mean, m and sample variance, s^2 are independent.

1.3 Distribution of $F(X)$

Suppose X is a random variable with distribution function $F(x)$, then $U = F(X)$ is a uniform random variable on [0,1]. To see this note that $F(\cdot)$ is strictly increasing and then $\Pr(U \leq u) = \Pr(F(X) \leq u) = \Pr(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u$.

1.4 Multinomial distribution

Suppose n items are classified as being in one of k categories, with the number in category j being the random variable R_j. If each item is classified independently of the other items and the probability of being in category j is π_j (with $\pi_1 + \ldots + \pi_k = 1$), then the $(R_1, \ldots, R_k)$ follow the *Multinomial distribution*, with

$$\Pr(R_1 = r_1, \ldots, R_k = r_k) = \frac{n!}{r_1!r_2!\ldots r_k!}\pi_1^{r_1}\pi_2^{r_2}\ldots\pi_k^{r_k}$$

where $r_1 + \ldots + r_k = n$. Note that $k = 2$ is the familiar Binomial distribution.

1.5 Covariances

This is trivial but perhaps worth a reminder so that some of the following derivations can be done more smoothly. The covariance of random variables X, Y is $\mathsf{cov}[X, Y] = \mathsf{E}[(X - \mathsf{E}[X])(Y - \mathsf{E}[Y])] = \mathsf{E}[XY] - \mathsf{E}[X]\mathsf{E}[Y]$. Consequently, if either $\mathsf{E}[X]$ or $\mathsf{E}[Y]$ vanishes, then $\mathsf{cov}[X, Y] = \mathsf{E}[XY]$.

1.6 Basu's Theorem

Suppose data are observed from a model $f(\cdot \mid \theta)$ and that i) T is complete and sufficient statistic for θ and ii) V is a statistic which has a distribution which does not depend on θ, then T and V are independent.

2 Normal probability plots

2.1 Rationale and method

An obvious way to assess the Normality of a sample is to draw a histogram and judge whether the shape looks reasonably close to the bell-shaped curve you would expect. What this is essentially doing is to see how many observations are in the lower and upper tails and how many are close to the central part of the sample. Put this way it is unsurprising that a better, but still essentially subjective, alternative is based on putting the sample into order.

To start with, suppose we have a sample of n independent observations from a standard Normal distribution, i.e a $N(0,1)$ distribution, namely $Z_1, Z_2, \ldots, Z_n$. Reorder these values to obtain $Z_{(1)} < Z_{(2)} < \ldots < Z_{(n)}$[1], where parentheses on the subscripts indicate that these are the *ordered* sample values, sometimes called the *order statistics*, with $Z_{(1)}$ being the smallest of the Z_i, $Z_{(2)}$ the second smallest, and so on. In Figure 1, the ordered values for samples of sizes 5, 10, 25, 50, 75 and 100 are shown as black circles. In a random sample, there will be considerable variation in the locations of individual points. What we can compute are the *expected order statistics*, $\mathsf{E}[Z_{(i)}]$, which are shown as red crosses in Figure 1. How these expectations are evaluated will be considered in Section 3.

From Figure 1 we see that the expected order statistics have the following properties:

1. they are symmetrically distributed about the centre of the sample;

2. there are some observations in the tails, but the bulk are concentrated in the centre, with the density being maximal near the mean and reducing towards the tails;

3. their range increases with sample size.

The random samples behave in a broadly similar way, albeit with inevitable random variation. The idea of a Normal Probability Plot (NPP) is to plot the ordered sample values against the expected order statistics from a standard Normal variable. If the data are from a standard Normal distribution, it would be expected that the plot would be a straight line, but for random deviations. This is shown in Figure 2(a) for the sample of size 50 in Figure 1.

If the data, $Y_1, Y_2, \ldots, Y_n$ are a random sample from a Normal distribution with mean μ and variance σ^2, then we can write $Y_i = \mu + \sigma Z_i$, where the Z_i are standard Normal. As $\sigma > 0$, it follows that $Y_{(i)} = \mu + \sigma Z_{(i)}$, so a plot of

[1] In this article attention is restricted to continuous distributions, so it is immaterial whether we use $<$ or $\leq$, as the two differ only by events with probability 0

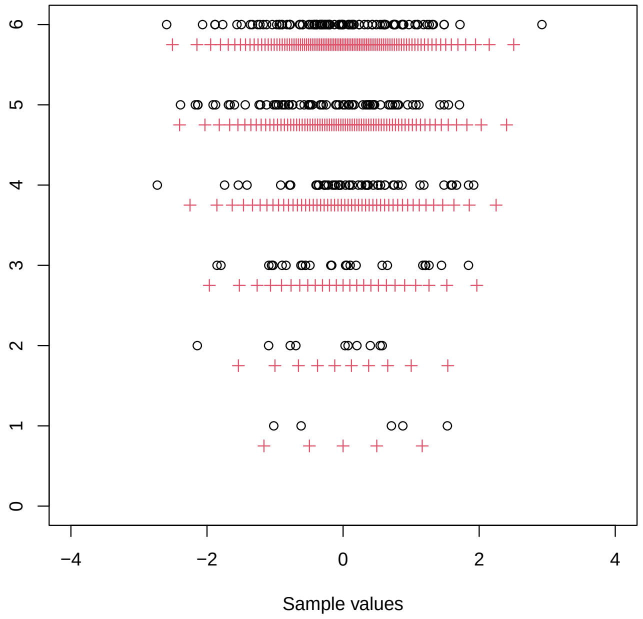
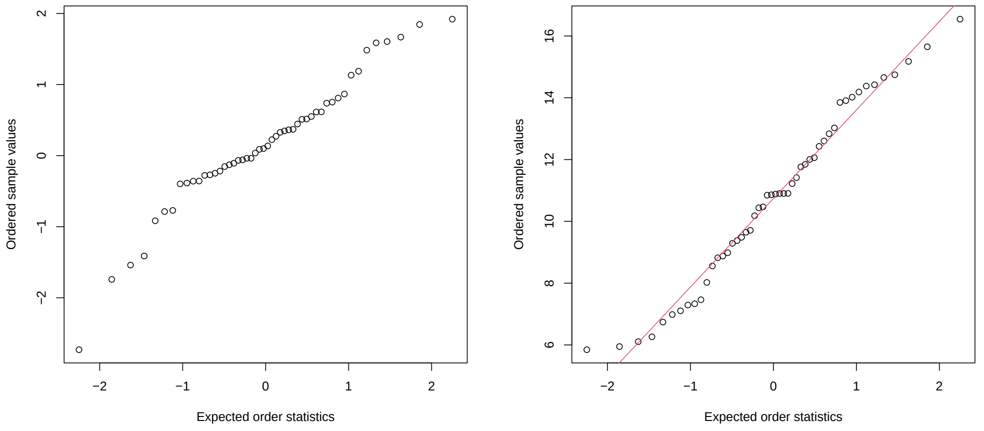


Figure 1: Plots of samples from the standard Normal distribution of sizes 5, 10, 25, 50, 75 and 100, black circles from bottom to top. Values of the expected order statistics for samples of the same size are the red crosses below each sample

the ordered data against the expected order statistics of a standard Normal distribution will still approximate a straight line, but now with intercept μ and slope σ. This is illustrated in Figure 2(b), where the 50 data points are from a Normal distribution with mean 10 and standard deviation 3. A simple regression, as performed in R using `lm()`, gives an intercept of 10.74 and slope of 2.87, broadly in line with the parameter values of 10 and 3, respectively.

2.2 Effectiveness of Normal plots

While NPPs provide a line that should be straight if the data are Normal, there will inevitably be departures from strict linearity. It therefore becomes a matter of judgment whether a NPP provides evidence that data can reasonably be assumed to be Normal. In this respect it is helpful to consider



(a) Sample size 50

(b) Sample from $N(10, 9)$ with $n = 50$

Figure 2: Plots of ordered observations against expected order statistics from standard Normal variable. Plot 2(a) is for sample of size 50 from Figure 1. Plot 2(b) is a new sample from $N(10, 9)$: red line is simple regression fit.

how far from linear is the NPP for data known to be Normal and for data known to be non-Normal.

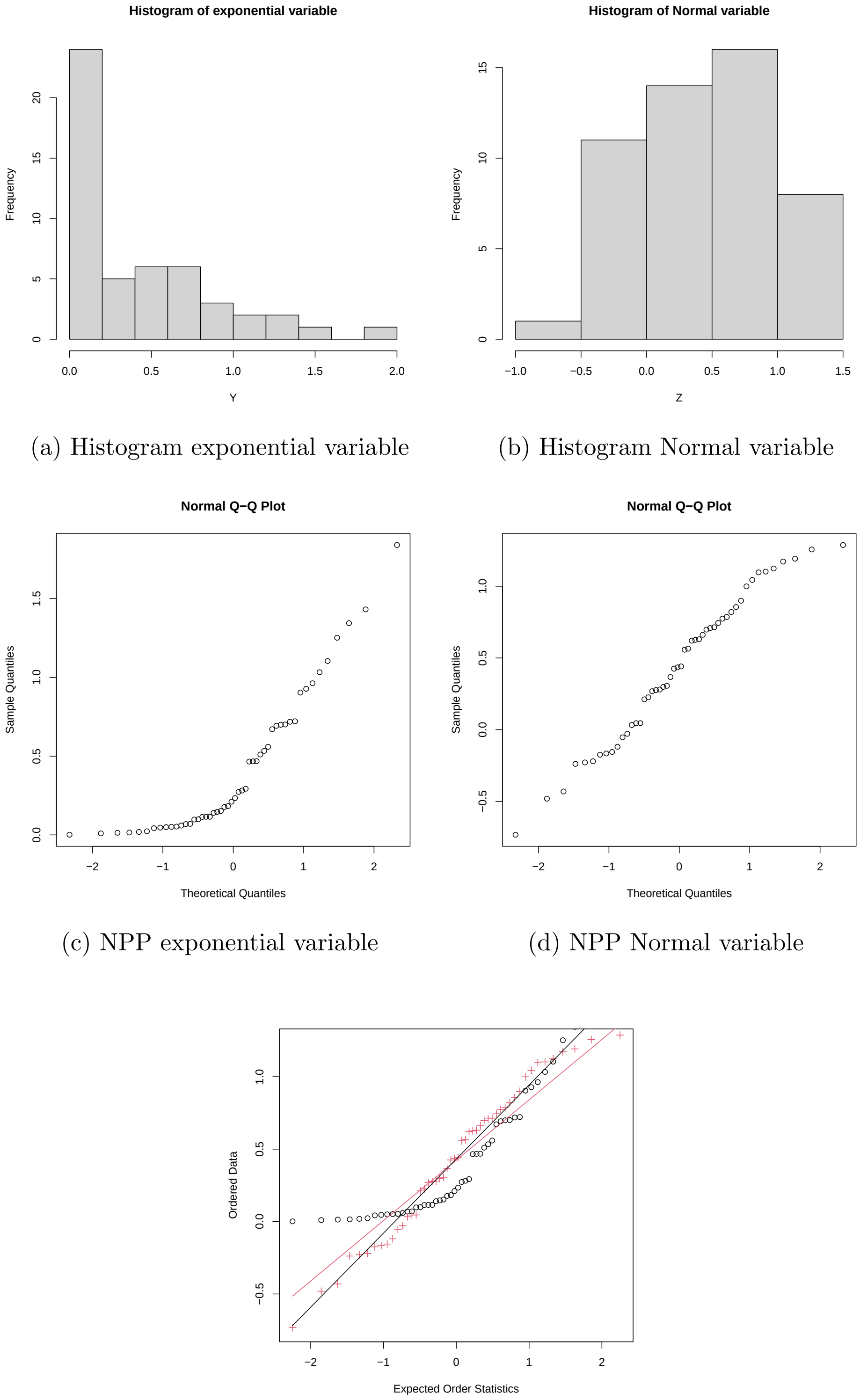
Of course, data can be non-Normal in many ways and it is useful to draw some distinctions. A Normal random variable, X, has zero *skewness*, as $\mathsf{E}[(X - \mu)^3]/\sigma^3 = 0$, and *kurtosis* 3, as $\mathsf{E}[(X - \mu)^4]/\sigma^4 = 3$.

2.3 Distinguishing Normal samples from skewed samples

Take as an example of a skewed random variable, Y, an exponential random variable with mean $\frac{1}{2}$, which has standard deviation of $\frac{1}{2}$. Figure 3 shows histograms and NPPs for a random sample of size 50 from the exponential distribution and, for comparison, a sample of size 50 from a Normal distribution with the same mean and standard deviation.

The histograms use the default number of bins from the R command `hist()`. It might be argued that the number of bins might profitably have been larger in Figure 3(b), although this also illustrates that no such choice is required when using NPPs.

Some care is needed in interpreting the NPPs in Figures 3(c) and 3(d). For skewed data, as shown by the exponential variable, there are a large number of observations near zero - these do not spread out as much as they would from the lower tail of a symmetric distribution, which leads to the relatively flat part of the NPP for the smaller abscissae. At the larger abscissae the ordinates increase rapidly. In Figure 3(d) the slope of the plot is fairly constant. However, it must be conceded that the immediate visual impression of Figures 3(c) and 3(d) is, perhaps, less striking than between the histograms. However, careful inspection of Figure 3(c) shows the clear curvature in the plot. This is seen very clearly in Figure 3(e), where simple fitted regression lines help guide the eye, and help to contrast the two NPPs more clearly.



(a) Histogram exponential variable

(b) Histogram Normal variable

(c) NPP exponential variable

(d) NPP Normal variable

(e) Normal (red) and exponential (black) variables

Figure 3: Histograms and NPPs for exponential and Normal variables, each with mean and SD equal to $\frac{1}{2}$. Figure 3(e) shows both on same axes, with simple fitted lines, with red for Normal and black for exponential.

2.4 Identifying non-Normal samples on the basis of kurtosis

While symmetry is clearly a necessary condition for a distribution to be Normal, there are symmetric distributions that are not Normal. In these cases perhaps the principal departure is through the heaviness of the tails, and this can be quantified by the kurtosis of the distribution. The *kurtosis* of a random variable, X, is defined as

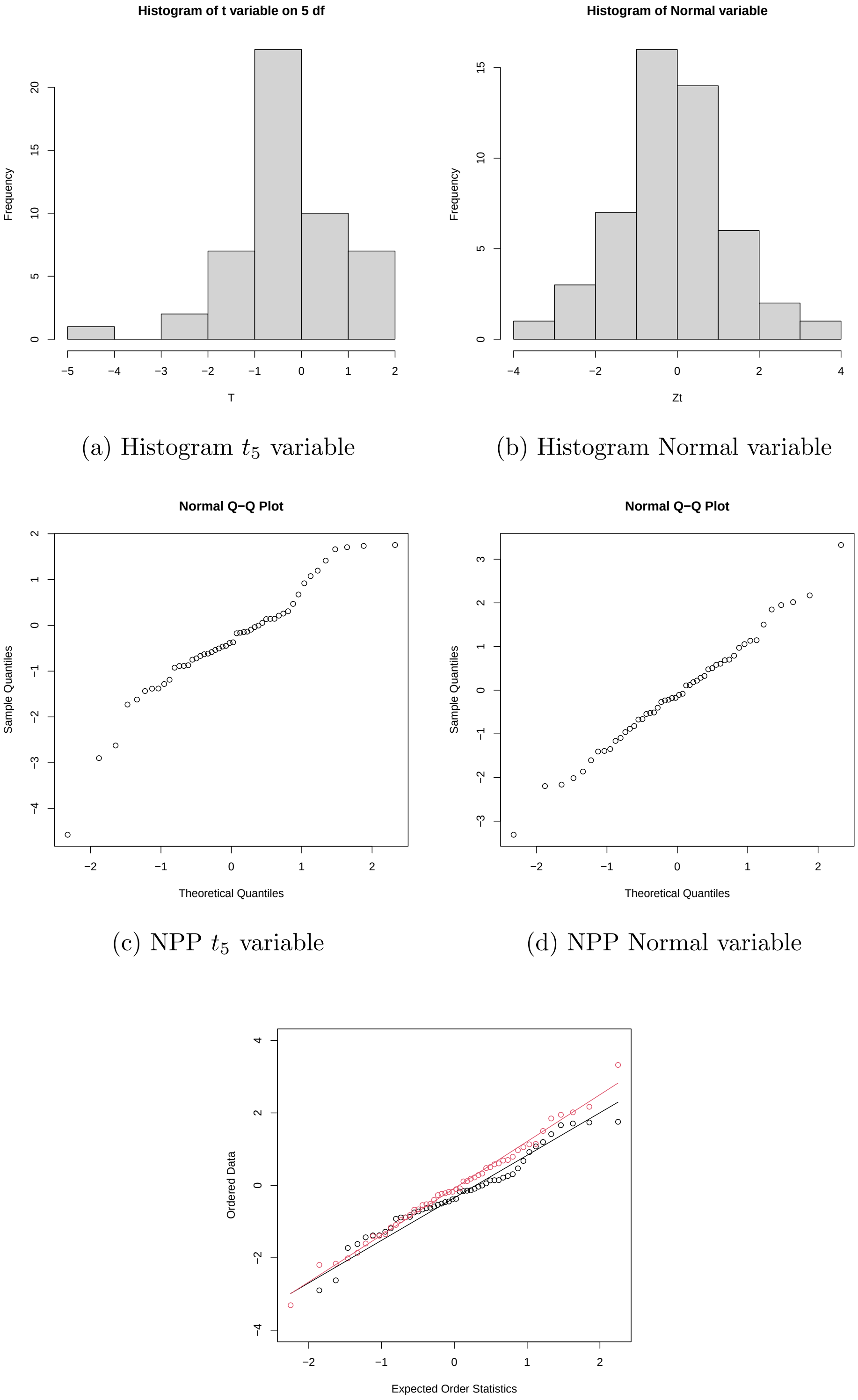
$$\kappa = \frac{\mathsf{E}[(X-\mu)^4]}{\sigma^4},$$

where μ, σ^2 are, respectively, the mean and variance of X. Note that κ is dimensionless and for a Normal distribution $\kappa = 3$. This leads to the definition of *excess* kurtosis as $\kappa - 3$. Distributions with $\kappa < 3$ are *platykurtic* and have lighter tails than the Normal, whereas heavier-tailed distributions are *leptokurtic* and have $\kappa > 3$.

An example of a leptokurtic distribution is the t-distribution on ν degrees of freedom. For large ν, t_ν is very similar to the Normal distribution, so the most notable differences are shown by taking ν to be small. However, in the above it has been assumed that all necessary moments, i.e. up to the fourth, exist but for the t-distribution this is not true for $\nu \leq 4$. For $\nu > 4$ the excess kurtosis is $6/(\nu - 4)$, so we use as an example $\nu = 5$, the smallest integer degrees of freedom for which the kurtosis exists, and where it is equal to 9 (excess kurtosis 6).

As might be imagined, it is hard to distinguish non-Normality on the basis of kurtosis alone. Because the differences arise in the tails, this will be especially true for small samples. In Figure 4(e) it is hard to distinguish the Normal from the t-sample. In Figure 4(c), the tails do seem to have larger values than might be expected for a Normal distribution, but the differences are more subtle than in Figure 3.

Of course, it may well be that the untoward effects on the analysis of non-zero excess kurtosis may be less marked than those of non-zero skewness, so detection of non-Normal kurtosis may be less important.



(a) Histogram t_5 variable

(b) Histogram Normal variable

(c) NPP t_5 variable

(d) NPP Normal variable

(e) Normal (red) and t_5 (black) variables

Figure 4: Histograms and NPPs for t_5 and Normal variables, each with mean 0 and variance 5/3. Figure 4(e) shows both on same axes, with simple fitted lines, with red for Normal and black for t_5.

3 Order statistics

A NPP plots the ordered data against the expected order statistics from a sample of the same size, drawn from a standard Normal distribution. How are these expected order statistics calculated? In this section we will consider this problem and we will start by deriving the distribution of the order statistics for a general univariate distribution. The next step, computing the expectations, is straightforward for a uniform distribution and explicit results are available for the exponential distribution. However, for the Normal distribution, explicit expectations are unavailable (except for a few special and unimportant results when the sample size is very small), so expectations have to be derived numerically, or through approximations. We restrict consideration to continuous random variables.

3.1 Distribution of order statistics for a general continuous distribution.

Suppose $X_1, \ldots, X_n$ is a sample of independent random variables, each with density $f(x)$ and distribution function $F(x)$. The distribution of elements of the ordered sample, $X_{(1)}, \ldots, X_{(n)}$ will differ, e.g. the distribution of $X_{(1)}$ is not the same as that of X_1 - it is intuitively obvious that $\mathsf{E}(X_{(1)})$ will be less than $\mathsf{E}(X_1)$ (unless, of course, $n = 1$). There are two slightly different ways to derive the distribution of $X_{(r)}$, for $r = 1, \ldots, n$. One calculates the density, $f_r(x)$ directly, whereas the other approach, adopted here, first finds the distribution function, $F_r(x)$, and derives the density by differentiation. The latter is conceptually slightly simpler, albeit finishing with a longer, although routine computation.

We have $F_r(x) = \Pr(X_{(r)} < x)$, and the event $X_{(r)} \leq x$ will occur if *at least* r of $X_1, \ldots, X_n$ are less then x, and of course the chance any element of the unordered sample is less than or equal to x is $F(x)$. Consequently $\Pr(X_{(r)} < x)$ is the probability of at least r successes in n binomial trials, where the probability of success is $F(x)$. Hence

$$F_r(x) = \sum_{j=r}^{n} a_j \tag{1}$$

where

$$a_j = \binom{n}{j} F(x)^j [1 - F(x)]^{n-j}, \qquad j = 1, \ldots, n.$$

To obtain the density of $X_{(r)}$, $f_r(x)$, (1) must be differentiated with respect to x. For $r \leq j \leq n$

$$\begin{aligned}\frac{da_j}{dx} &= j \binom{n}{j} F(x)^{j-1} [1 - F(x)]^{n-j} f(x) - (n-j) \binom{n}{j} F(x)^j [1 - F(x)]^{n-j-1} f(x) \\ &= b_j - c_j, \qquad \text{say.}\end{aligned}$$

Using the observation that

$$(j+1) \binom{n}{j+1} = \frac{n!}{j!(n-j-1)!} = (n-j) \binom{n}{j},$$

it follows that $b_{j+1} = c_j$, so consequently the sum of the $b_j - c_j$ telescopes and hence

$$f_r(x) = b_r - c_n = r \binom{n}{r} F(x)^{r-1} [1 - F(x)]^{n-r} f(x). \tag{2}$$

The multiplier has an alternative form because

$$n \binom{n-1}{r-1} = r \binom{n}{r}.$$

The alternative derivation leads to this other multiplier. Briefly, the argument is that the probability $X_{(r)}$ is in $(x, x + dx)$ is found by assuming X_1 is in $(x, x + dx)$ and that $r - 1$ of the remaining $n - 1$ variables are less than x. This needs to be multiplied by n because any of the X_j could be the one in $(x, x + dx)$.

3.2 Order statistics for uniform and Normal distributions

3.2.1 Uniform distribution

If the X_j are uniformly distributed on $[0, 1]$, then $F(x) = x$ and $f(x) = 1$. Consequently, the rth order statistic from a sample of n independent

Uniform[0,1] variables has density

$$n \binom{n-1}{r-1} x^{r-1}(1-x)^{n-r},$$

i.e. it has a Beta$(r, n-r+1)$ distribution, which has mean $r/(n+1)$ and variance $r(n-r+1)/[(n+1)^2(n+2)]$.

3.2.2 Normal distribution

For the Normal distribution, substituting $F = \Phi$ and $f = \phi$ in (2) does not yield a distribution that is tractable, at least in the sense of explicit expressions for the means and variances of the order statistics. It is always possible to compute $\mathsf{E}[X_{(r)}]$ by numerical evaluation of $\int x f_r(x) dx$, as in Royston (1982b), which is implemented in the function `evNormOrdStats` in the R library `EnvStats`: see also the amendment in Königer (1983).

However, this is computationally demanding, especially for larger n and a good and easily computed approximation would be useful. This can be obtained applying the result in 1.3. Suppose $Z_1, \ldots, Z_n$ are independent values from a standard Normal distribution, then $\Phi(Z_1), \ldots, \Phi(Z_n)$ is a sample from a Uniform distribution on [0,1]. As Φ is increasing, $\Phi(Z_{(r)})$ is the rth order statistic from the sample $\Phi(Z_1), \ldots, \Phi(Z_n)$, so $\mathsf{E}[\Phi(Z_{(r)})] = \mathsf{E}[U_{(r)}] = r/(n+1)$.

As a first attempt at an approximation, we take $\mathsf{E}[\Phi(Z_{(r)})] \approx \Phi(\mathsf{E}[Z_{(r)}])$, and hence

$$\mathsf{E}[Z_{(r)}] \approx \Phi^{-1}(r/(n+1)). \tag{3}$$

Interchanging $\mathsf{E}[\cdot]$ and $\Phi(\cdot)$ is entirely bogus but works quite well. Blom (1958) explored a range of other approximations, including the family of approximations

$$\mathsf{E}[Z_{(r)}] \approx \Phi^{-1}\left(\frac{r-\alpha}{n-2\alpha+1}\right),$$

for $0 \leq \alpha \leq 1$. Blom found the best approximation was to use $\alpha = \frac{3}{8}$, i.e.

$$\mathsf{E}[Z_{(r)}] \approx \Phi^{-1}\left(\frac{r-\frac{3}{8}}{n+\frac{1}{4}}\right). \tag{4}$$

For the case $n = 50$ and for $r = 40, \ldots, 50$, $\mathsf{E}[Z_{(r)}]$ are shown in Table 1 for the method of Royston and using (3) and (4), and also $\alpha = \frac{1}{2}$, as this is a version used in half-normal plots - see Section 5.2. It can be seen that Blom's approximation is much better that the first attempt (3) and is better than that using $\alpha = \frac{1}{2}$. Only the largest 11 order statistics have been shown. The table for $r = 1, \ldots, 11$ is the same but for a change of sign. The approximations are all poorest in the tails. Blom's approximation remains the best for all r.

r	exact	$\alpha = 0$	$\alpha = \frac{3}{8}$	$\alpha = \frac{1}{2}$
40	0.8023	0.7868	0.8014	0.8064
41	0.8732	0.8557	0.8722	0.8779
42	0.9489	0.9289	0.9477	0.9542
43	1.0304	1.0074	1.0290	1.0364
44	1.1195	1.0927	1.1177	1.1264
45	1.2185	1.1868	1.2163	1.2265
46	1.3311	1.2928	1.3283	1.3408
47	1.4637	1.4157	1.4600	1.4758
48	1.6286	1.5647	1.6235	1.6449
49	1.8549	1.7599	1.8475	1.8808
50	2.2491	2.0619	2.2433	2.3263

Table 1: Exact and approximate values of $\mathsf{E}[Z_{(r)}]$ for $n = 50$, showing $r = 40, \ldots, 50$: $\alpha = 0$ corresponds to (3), $\alpha = \frac{3}{8}$ to Blom's approximation (4) and the final column uses $\alpha = \frac{1}{2}$, an approximation that is sometimes encountered.

3.3 Dependence of order statistics

It is obvious that the order statistics, $X_{(1)}, \ldots, X_{(n)}$ are dependent, even when the underlying sample values $X_1, \ldots, X_n$ are independent. To see this note that while $\Pr(X_1 < x \mid X_2 < x) = \Pr(X_1 < x)$, it is clear that $\Pr(X_{(1)} < x \mid X_{(2)} < x) = 1 \neq \Pr(X_{(1)} < x)$, as the smallest sample value must be less than the second smallest. Elucidating more quantitative information requires straightforward but intricate analysis, details of which are in Appendix A. From the Appendix, the joint distribution of $X_{(r)}$ and $X_{(s)}$, with $r < s$, $f_{rs}(x, y)$ can be written as $f_{rs}(x, y) = 0$ for $x > y$ and for

$x < y$

$$\begin{aligned} f_{rs}(x,y) = {} & \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \\ & \times F(x)^{r-1}[F(y)-F(x)]^{s-r-1}[1-F(y)]^{n-s}f(x)f(y). \end{aligned} \tag{5}$$

This expression is not that important in itself but it is needed to evaluate the $n \times n$ dispersion matrix of the order statistics. This is used in the Shapiro-Wilk test (see Section 4.1), although it turns out that even there the role of the covariances is not that critical, and may even be inappropriate.

4 Testing for Normality: Shapiro-Wilk and Shapiro-Francia

If we are uncertain whether or not a sample is Normally distributed, perhaps we should test the null hypothesis that the sample comes from a Normal distribution? This is certainly possible and can be done by several methods. Perhaps the simplest are methods to test if the sample skewness and kurtosis are compatible with the values of 0 and 3, respectively. A more sophisticated method is that proposed by Shapiro & Wilk (SW) (Shapiro and Wilk, 1965), and its slightly simpler cousin due to Shapiro & Francia (SF) (Shapiro and Francia, 1972), and these will be introduced and explained in this Section. There are, of course, many ways to test for Normality, and there are whole books on the subject, see, e.g., Thode (2002). We focus on SW and SF because they turn out to have good properties and, compared with approaches such as testing skewness and kurtosis, the way they test for Normality probably needs more explanation.

Before doing so it worth pausing to think whether the use of such tests is likely to be helpful, and some issues to bear in mind are below.

- Some tests, such as a test for skewness, will focus on one aspect, and even if there is no evidence of non-Normal skewness, there could be other aspects that mean the data are not Normal. Also, the tests may lack power and non-significant results may be uninformative.

- Other tests, such as the SW and SF tests, are focussed on the very general alternative that the data are not Normally distributed. Here power can be even more of a problem, although such tests do avoid focussing on one aspect of Normality.

- Perhaps the most important thing to keep in mind is why you want to assess data for Normality. If you simply want some reassurance that the data are broadly in conformity with the assumptions made in your model, then you may well be able to accept some departure from Normality. Many statistical models are very forgiving of modest departures from Normality.

- More care is needed if the analysis puts a heavier reliance of the assumption of Normality. One example of this is when using data to compute reference ranges, for example age-related quantiles of height for children. It is much more efficient to estimate the third centile of a Normal population by $m - 1.88s$ than by using some version of $X_{(0.03n)}$. However, the assumption of Normality is now of much more importance than when assessing the fit of a regression model.

- Although not very helpful advice, assessment of Normality is probably akin to Whistler's comment[2].

4.1 The Shapiro-Wilk test

As we have observed elsewhere, if $Y_1, \ldots, Y_n$ is a sample from a Normal distribution with unknown mean μ and unknown variance σ^2, then $Y_{(i)}$ can be written as $\mu + \sigma Z_{(i)}$, where $Z_1, \ldots, Z_n$ are independent random variables from a standard Normal distribution. It will be convenient to write $m_i = \mathsf{E}[Z_{(i)}]$, with $\boldsymbol{m}$ being the vector of the m_i, and $\boldsymbol{V}$ for the dispersion matrix of the $Z_{(i)}$. As $\mathsf{E}[Y_{(i)}] = \mu + \sigma m_i$, a regression of the ordered sample values on $\boldsymbol{m}$ will yield an intercept that estimates μ and a slope that estimates σ. The latter is based on the assumption that the data are Normal. The usual estimate of σ^2 is valid for any distribution and it is the comparison of these two estimates that is the basis of the SW test.

Assuming the Y_i are Normal, then the $Y_{(i)}$ have dispersion $\sigma^2\boldsymbol{V}$, so the most efficient estimator of $(\mu, \sigma)^T$ is a generalised least squares regression, with weight matrix $\boldsymbol{V}^{-1}$. If $\mathbf{1}$ denotes an n-dim vector of ones, then

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix} = \begin{pmatrix} \mathbf{1}^T\boldsymbol{V}^{-1}\mathbf{1} & \mathbf{1}^T\boldsymbol{V}^{-1}\boldsymbol{m} \\ \boldsymbol{m}\boldsymbol{V}^{-1}\mathbf{1} & \boldsymbol{m}^T\boldsymbol{V}^{-1}\boldsymbol{m} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}^T\boldsymbol{V}^{-1}\boldsymbol{y} \\ \boldsymbol{m}^T\boldsymbol{V}^{-1}\boldsymbol{y} \end{pmatrix} \tag{6}$$

where $\boldsymbol{y}$ is the vector of the ordered values of the observed sample. Considerable simplification is possible because $\boldsymbol{V}^{-1}\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^T\boldsymbol{m} = 0$. The latter would be true for any symmetric distribution but the former requires the observations to be Normal. These results, and some others of use below are derived in Appendix B. Consequently we have $\hat{\mu} = \overline{y}$ and $\hat{\sigma} = \boldsymbol{m}^T\boldsymbol{V}^{-1}\boldsymbol{y}/(\boldsymbol{m}^T\boldsymbol{V}^{-1}\boldsymbol{m})$.

[2] James Whistler (1834-1903) was an American-born artist. A painting that had taken him two days to complete cost 200 guineas. When asked if the work of two days was worth 200 guineas, he replied it was for the work of two days and the experience of a lifetime.

Shapiro & Wilk defined their statistic, W, as essentially the ratio of the square of this estimator of σ to the usual estimator of σ^2, namely s^2. In fact, Shapiro and Wilk chose the numerator to be $(\boldsymbol{a}^T\boldsymbol{y})^2$, where $\boldsymbol{a}$ is a unit vector that is proportional to $\boldsymbol{V}^{-1}\boldsymbol{m}$, and the denominator as $(n-1)s^2$. As s^2 is independent of W (another application of Basu's Theorem), we have the null expectation of W to be

$$\mathsf{E}[W] = \frac{\mathsf{E}[(\boldsymbol{a}^T\boldsymbol{y})^2]}{\mathsf{E}[(n-1)s^2]} = \frac{\boldsymbol{m}^T\boldsymbol{V}^{-1}\boldsymbol{m}(1+\boldsymbol{m}^T\boldsymbol{V}^{-1}\boldsymbol{m})}{(n-1)\boldsymbol{m}^T\boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{m}}.$$

These quadratic forms, being based on the standard Normal variable, can be calculated. A simulation of a million samples of size 50 provided an estimate of $\boldsymbol{V}$, while $\boldsymbol{m}$ can be found from `evNormOrdStats` in library `EnvStats` in R . To provide some feeling for these quantities, the values for $n = 50$ are

$$\mathsf{E}[W] = \frac{97.2048 \times 98.2048}{49 \times 201.3733} = 0.967.$$

The mean is inevitably less than 1 because an application of the Cauchy-Schwarz inequality shows that $W \leq 1$.

The null hypothesis is discredited when W is too small. An important contribution to determining the significance level of the test was made by Royston (1982a), who applied a Box-Cox transformation to $1 - W$ to facilitate the calculation of percentage points.

4.2 The Shapiro-Francia Test

The main difficulty with the SW test is computing $\boldsymbol{V}$, which is used because it is proportional to the dispersion matrix of the observations in the regression estimating μ and σ. However, using ordinary least squares, rather then generalized least squares, will still provide consistent estimators for μ and σ and, as $\boldsymbol{V}$ is no longer needed, will be considerably easier to calculate.

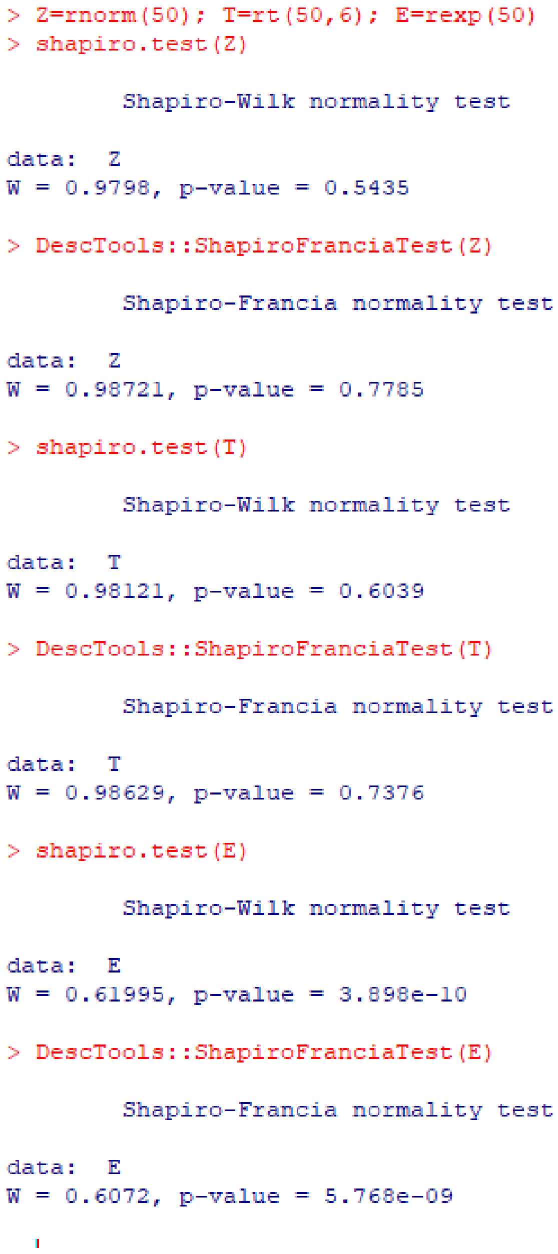
This is the approach suggested by Shapiro and Francia (1972) and the resulting test is the Shapiro-Francia (SF) test. The test statistic, W_f, has the same form as (6) but with $\boldsymbol{V}$ replaced by $\boldsymbol{I}$. So W_f uses $(\boldsymbol{b}^T\boldsymbol{y})^2$ in the numerator, where $\boldsymbol{b}$ is a unit vector proportional to $\boldsymbol{m}$, and $(n-1)s^2$ in the denominator. As with W, W_f is independent of s^2 so

$$\mathsf{E}[W_f] = \frac{\boldsymbol{m}^T\boldsymbol{V}\boldsymbol{m} + (\boldsymbol{m}^T\boldsymbol{m})^2}{(n-1)\boldsymbol{m}^T\boldsymbol{m}}.$$

For samples of size 50 $\boldsymbol{m}^T\boldsymbol{m} = 47.4217$, $\boldsymbol{m}^T\boldsymbol{V}\boldsymbol{m} = 23.1485$, and hence $\mathsf{E}[W_f] = 0.9778$. Again the Cauchy-Schwarz inequality indicates $W_f \leq 1$. For this test if $\boldsymbol{y} = \overline{y} + s\boldsymbol{m}$, i.e. if the observed sample is distributed exactly as its expectation, then $W_f = 1$, thus demonstrating that it is values in the lower tail of the distribution of W_f that are associated with non-Normality.

The properties of W_f and W are very close, so there is no compelling reason to use the complicated W. Indeed, W_f can be further simplified by replacing m_i in $\boldsymbol{b}$ with the approximation suggested by Blom in (4). This was first proposed by Weisberg and Bingham (1975), who confirmed that using these values gives a test statistic that is almost indistinguishable from W_f.

4.3 The tests in practice



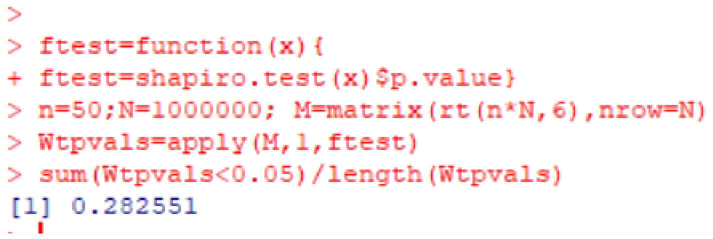
```
> Z=rnorm(50); T=rt(50,6); E=rexp(50)
> shapiro.test(Z)

        Shapiro-Wilk normality test

data:  Z
W = 0.9798, p-value = 0.5435

> DescTools::ShapiroFranciaTest(Z)

        Shapiro-Francia normality test

data:  Z
W = 0.98721, p-value = 0.7785

> shapiro.test(T)

        Shapiro-Wilk normality test

data:  T
W = 0.98121, p-value = 0.6039

> DescTools::ShapiroFranciaTest(T)

        Shapiro-Francia normality test

data:  T
W = 0.98629, p-value = 0.7376

> shapiro.test(E)

        Shapiro-Wilk normality test

data:  E
W = 0.61995, p-value = 3.898e-10

> DescTools::ShapiroFranciaTest(E)

        Shapiro-Francia normality test

data:  E
W = 0.6072, p-value = 5.768e-09
```

On the left is a screenshot of the application of the SW and SF tests in R: the former from the base function `shapiro.test` and the latter using `ShapiroFranciaTest` from the `DescTools` package.

The tests have been applied to three samples, each of size 50. The variable `Z` is from a Normal populations, `T` is from a *t*-distribution with 6 df, i.e. a highly leptokurtic distribution, and `E` is from an exponential distribution with unit mean, i.e. a highly skewed distribution.

As can be seen, p-values exceeding 0.05 were obtained, not only for the Normal sample but also for the *t*-sample. For the exponential sample, highly significant p-values were obtained. This illustrates that the tests have much less power against alternatives that have non-Normal kurtosis than non-Normal skewness - something we might anticipate from a histogram.

These functions are also useful in that they allow the user to assess rapidly the power of a test in a particular circumstance. For the example of samples of size 50 from a t-distribution on 6 df, the code below allows an estimate of the power of the SW test, which is about 0.3



```
>
> ftest=function(x){
+ ftest=shapiro.test(x)$p.value}
> n=50;N=1000000; M=matrix(rt(n*N,6),nrow=N)
> Wtpvals=apply(M,1,ftest)
> sum(Wtpvals<0.05)/length(Wtpvals)
[1] 0.282551
```

4.3.1 Examination of residuals

In a previous article, assessment of residuals was considered as a way to decide if the assumptions underpinning a model were reasonable. One common assumption, namely the Normality of the residuals, was deferred until this article, because a common method of assessment is to use a NPP of the estimated residuals. In addition, tests such as SW and SF can be applied to the residuals.

Consider the data on Ophthalmic Index (OI) and age used in the article on residuals. Fitting the models with either OI or $\log(OI)$ as the response variable gives rise to the test results in Table 2 when SW and SF are applied to the standardized residuals. The results from the two tests are very similar and both show that the model fitted to OI results in residuals that are not Normal, whereas there is no evidence of the residuals for the model using $\log(OI)$ departing from Normality.

	Shapiro Wilk	Shapiro Francia
OI	$W = 0.9543, p = 0.0015$	$W_f = 0.9545, p = 0.0024$
$\log(OI)$	$W = 0.9834, p = 0.235$	$W_f = 0.9837, p = 0.212$

Table 2: Shapiro-Wilk (W) and Shapiro-Francia (W_f) statistics and associated p-values applied to the standardized residuals from the regression of OI or $\log(OI)$ on age

The NPP plots applied to the standardized residuals from both models are shown in Figure 5. The NPP for residuals from the OI regression, black circles, appears to be curved above the line, whereas the NPP for the $\log(OI)$, red crosses, conforms more closely to the fitted line.

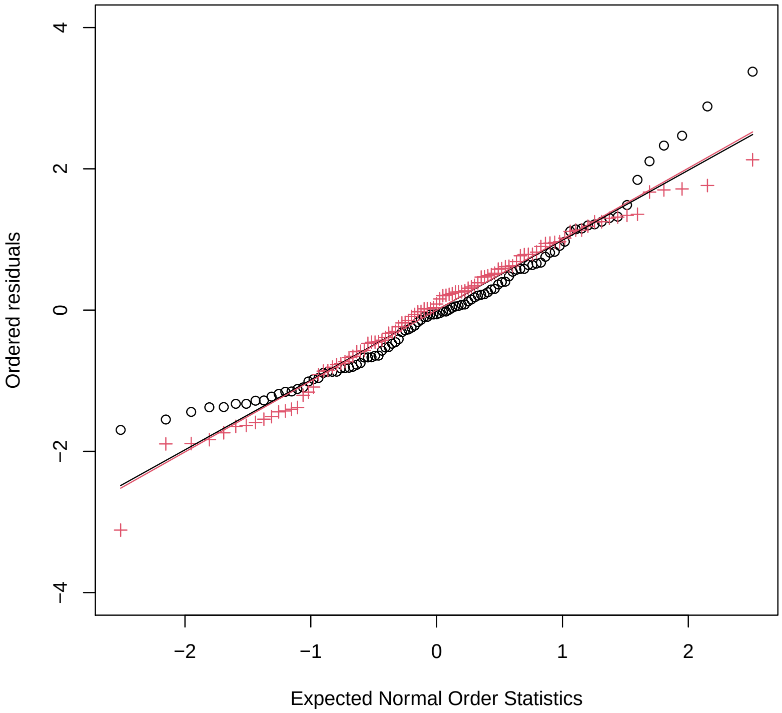


Figure 5: NPP of standardized residuals for response OI (black circles) and $\log(OI)$ (red crosses), with ordinary least squares fitted lines

4.3.2 SW versus SF tests when applied to residuals

The SF test was introduced as a numerically simpler test that was not much worse than the theoretically superior SW. The theoretical superiority of the SW test is based on its use of generalized least squares with weighting matrix $\boldsymbol{V}^{-1}$, where $\boldsymbol{V}$ is the dispersion matrix of the ordered values from a sample of *independent* standard Normal variables. However, when applied to residuals the original sample values are not independent.

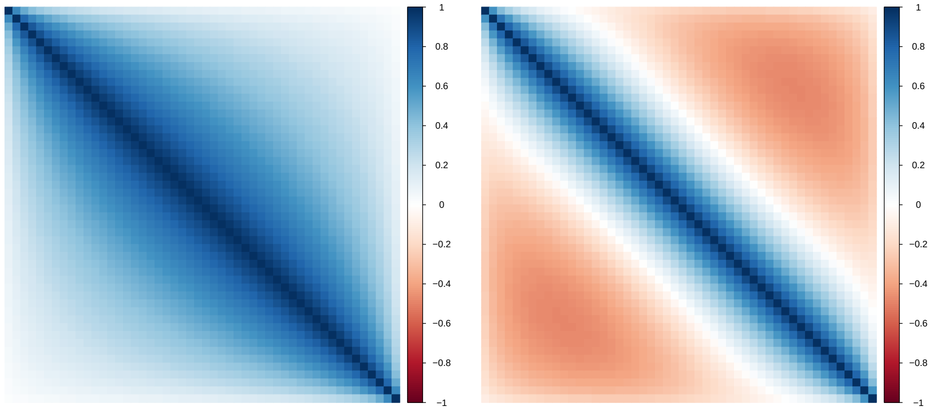
From an intuitive point of view, it might be thought that the dependence induced by ordering would be much stronger than that affecting residuals. The former affects each ordered value - e.g. the 3rd smallest value cannot be smaller than the 2nd smallest value, nor larger than the 4th smallest. On the other hand the dependence between residuals arises from the linear constraints to which they are subject: for the OI example there are 101 observations constrained by just two linear conditions. This can be seen in the correlations matrices in Table 3 for $n = 7$, a small sample size chosen so that the whole matrix can be presented. The residuals come from fitting a simple straight line (so the dependence of seven points with two constraints will exaggerate that seen in more realistic examples), and the correlations of the ordered values are found from simulating a million samples of size seven.

$$\begin{pmatrix} 1.00 & 0.62 & 0.45 & 0.34 & 0.26 & 0.19 & 0.11 \\ 0.62 & 1.00 & 0.73 & 0.56 & 0.43 & 0.31 & 0.19 \\ 0.45 & 0.73 & 1.00 & 0.77 & 0.59 & 0.43 & 0.26 \\ 0.34 & 0.56 & 0.77 & 1.00 & 0.77 & 0.56 & 0.34 \\ 0.26 & 0.43 & 0.59 & 0.77 & 1.00 & 0.73 & 0.45 \\ 0.19 & 0.31 & 0.43 & 0.56 & 0.73 & 1.00 & 0.62 \\ 0.11 & 0.19 & 0.26 & 0.34 & 0.45 & 0.62 & 1.00 \end{pmatrix}$$

$$\begin{pmatrix} 1.00 & -0.06 & -0.19 & -0.39 & -0.03 & -0.42 & 0.07 \\ -0.06 & 1.00 & -0.17 & -0.04 & -0.33 & -0.02 & -0.46 \\ -0.19 & -0.17 & 1.00 & -0.20 & -0.17 & -0.20 & -0.19 \\ -0.39 & -0.04 & -0.20 & 1.00 & 0.00 & -0.48 & 0.13 \\ -0.03 & -0.33 & -0.17 & 0.00 & 1.00 & 0.03 & -0.55 \\ -0.42 & -0.02 & -0.20 & -0.48 & 0.03 & 1.00 & 0.18 \\ 0.07 & -0.46 & -0.19 & 0.13 & -0.55 & 0.18 & 1.00 \end{pmatrix}$$

Table 3: Correlation matrices for ordered values of sample with $n = 7$ (upper matrix) and for residuals from simple regression (lower matrix).

Table 3 seems to bear out the intuition rehearsed in the previous paragraph: the correlations for the order statistics are positive and have larger magnitudes than for the residuals, whose correlations are also largely negative . As such, when applying the SW test to residuals one might expect the correlation induced by the ordering will be the main contributor to the dependence of the ordered residuals, so the use of weighting by $\boldsymbol{V}^{-1}$ may still be reasonable. However, using $n = 50$ and calculating the correlations of the ordered values of from a million simulations of i) independent standard Normal variables and ii) standardised residuals from a simple regression gives results shown in Figure 6.



(a) Independent values

(b) Residuals

Figure 6: Plot of correlation matrix of ordered independent observations (left) and ordered residuals (right)

Thus the dispersion matrices of the two forms of order statistics are quite different: the dispersion matrix shown in Figure 6(b) includes negative values, which are absent in Figure 6(a). As such, the use of $\boldsymbol{V}^{-1}$ in the computation of SW may be misplaced and assuming the more non-committal identity matrix used in the SF test may have theoretical as well as numerical advantages. This observation does not seem to appear in the literature and is only partial - e.g. the comparison in Figure 6 compares correlation not covariance matrices, but it is an issue worth keeping in mind.

A further comment on the assessment of the Normality of residuals that is, perhaps, worth making at this point, concerns the normality of $\boldsymbol{\epsilon}$, the unobserved residuals which the model assumes to be Normal, and the estimated residuals, $\hat{\boldsymbol{\epsilon}}$, which are the quantities we assess for Normality in lieu of the $\boldsymbol{\epsilon}$. Because $\hat{\boldsymbol{\epsilon}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon}$, the estimated residuals are linear combinations of the actual residuals. Consequently, even if $\boldsymbol{\epsilon}$ departed from Normality, the $\hat{\boldsymbol{\epsilon}}$ might be expected to depart less from Normality because of the Central Limit Theorem. This is a phenomenon sometimes referred to as *supernormality*, a term first coined by Gentleman and Wilk (1975) in the context of cross-classified data. In most circumstances this is something to be aware of, rather than worried by.

5 Some miscellaneous topics

In this final section a few topics related to the above are described briefly.

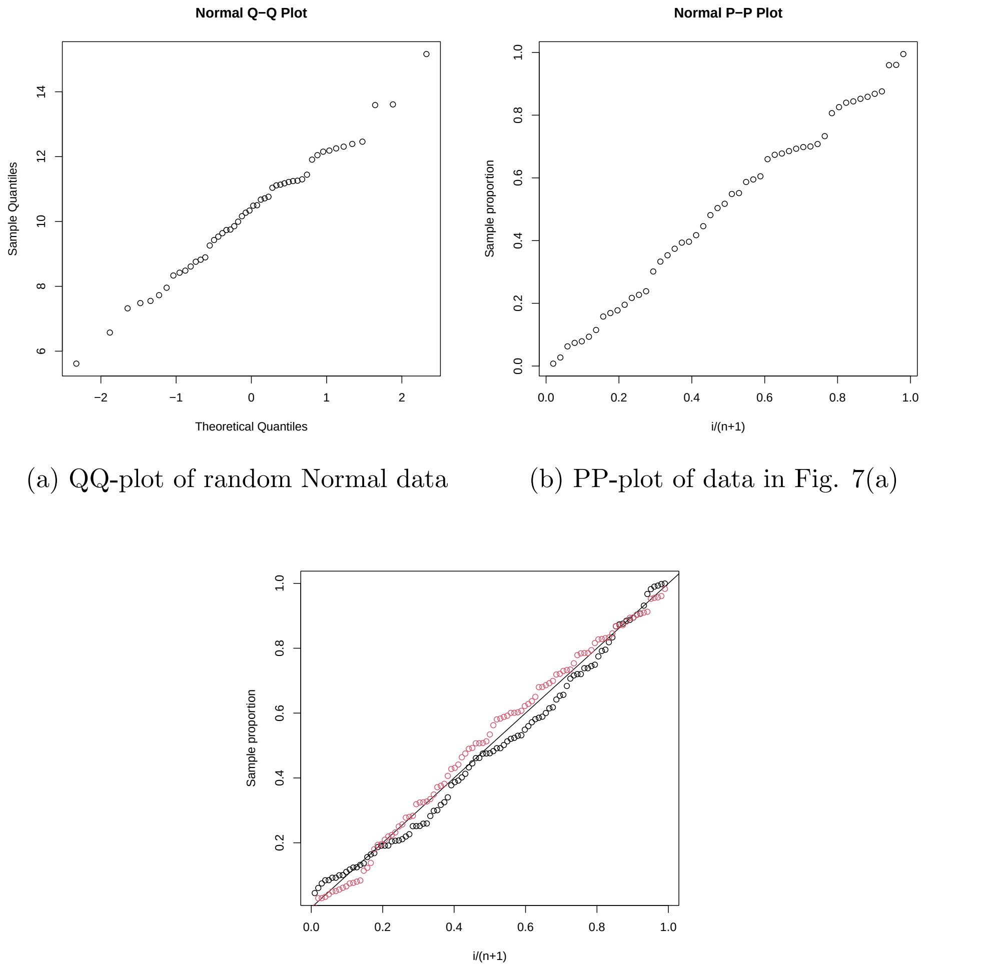
5.1 Q-Q plots and P-P plots

The NPP discussed above is an example of a Q-Q plot, short for quantile-quantile plot. In general a Q-Q plot plots the ordered values from the sample against the quantiles of a suitably chosen distribution. As it is the quantiles that are plotted, the range of the axes is determined by the support of the distribution in question. An approximate version is that the $Y_{(i)}$ are plotted against $F^{-1}(\mathsf{E}[U_{(i)}]) = F^{-1}(i/(n+1))$, for the appropriate distribution function $F(\cdot)$.
An alternative is to plot a P-P or probability-probability plot. Essentially this amounts to applying $F(\cdot)$ to both axes. The abscissa is straightforward, being $i/(n+1)$, while the ordinate needs some care to make the scale correct. For a Normal P-P plot, the ordinate is $\Phi((Y_{(i)} - \hat{\mu})/\hat{\sigma})$, i.e. unlike the Q-Q plot, parameter estimates are required. The range of both axes will be from 0 to 1 and if the data are Normal then the plot will vary about the line $y = x$.

Unlike a QQ-plot, where the best fit line needs to be determined, with a PP-plot the line the data should follow if the assumptions being tested are true is always $y = x$. Examples are shown in Figure 7. Note the scales in Figures 7(a) and 7(b) - while the plots are quite similar, it is perhaps hard to see departures in the tail with the PP-plot. The PP-plots of the residuals from the two models for the OI data are less easy to distinguish than is the case for the QQ-plots shown in Figure 5.

While the choice between PP and QQ plots is, to some degree, a matter of taste, most analysts probably prefer QQ-plots to assess goodness of fit.

Indeed, (Thode, 2002, p. 23) suggests that PP-plots are usually used when both axes plot samples, where the aim is to see if they come from the same population, rather than to assess if a sample has come from some prescribed distribution.



(a) QQ-plot of random Normal data

(b) PP-plot of data in Fig. 7(a)

(c) PP-plots from OI regression

Figure 7: QQ and PP plots of same random Normal sample (a) and (b): PP-plots of standardized residuals for regression of OI (black) and $\log(OI)$ (red) on age (c).

5.2 Half-normal plots

Half-normal plots (HNPs) are QQ-plots based on the half-normal distribution. The half-normal distribution is related to the folded normal distribution (Johnson et al., 1994, p. 170), the truncated normal distrbution (Johnson et al., 1994, p. 156) and is an extreme form of the Skew-normal distribution (Azzalini and Capitano, 2014). For our purposes we will take the half-normal distribution to be the distribution of the modulus of a Normal distribution with zero mean, $N(0, \sigma^2)$. In which case a random variable with this distribution is essentially $\sigma|Z|$, where Z is a standard Normal variable. The HNP will then plot the data against the order statistics, $W_{(i)}$ of a sample of $W_i = |Z_i|$.

The abscissae in a HNP of a sample of size n can be found using the approximation $F^{-1}(U_{(i)})$ discussed in Section 3.2.2. For the half normal distribution $F(x) = 2\Phi(x) - 1$ (remember $x > 0$) and the approximate expected order statistics, sometimes called half-normal scores, are $\Phi^{-1}(\frac{1}{2}(x_i + 1))$. A possibility is to set $x_i = i/(n+1), i = 1, \ldots, n$ but it is more usual to use $x_i = (i - \frac{1}{2})/n$, as used by Hills (1969) and in the R function `halfnorm` in the package `daewr`. Refinement of the approximation to the expected order statistics of the half-normal distribution seems to have received less attention than for the Normal distribution.

HNPs are used to assess Normality of variables expected to have zero mean and are effective at highlighting individual values that may not conform to this assumption amid data which generally seems to have zero mean. Early use of HNPs was in the interpretation of high-order factorial experiments (Daniel, 1959), where estimates for many main and interaction effects are produced and where most of the population effects are expected to be zero, so attention is on identifying the non-zero effects. Formal hypothesis testing will quickly run into problems of multiplicity and, in any case, a more informal assessment is usually needed to identify cases worthy of further investigation.

A similar application, due to Hills (1969), is the identification of genuine correlations from a large correlation matrix, although writing in 1969 what was considered large probably differs somewhat from current perceptions. Hills first transformed the correlations to Normality, using $z_{ij} = \tanh^{-1} r_{ij}$, and then considered a HNP of the $|z_{ij}|$. Points departing from the expected line $y = \sigma x$ identify correlations worthy of further study - see Hills's paper for an interesting example.

HNPs can be used with residuals, as they have expected values of zero, with the absolute value of the residual being plotted against the half-normal scores. There are suggestions in Atkinson (1981) and Atkinson (1982) that HNPs are preferred to NPPs, especially when used to identify outlying or influential points. This opinion was later softened slightly, (Atkinson, 1985, p. 36), with NPPs seen to be more informative for larger samples, say $n > 100$. HNPs for the standardized residuals from the regressions of OI and $\log(OI)$ on age are in Figure 8, where it seems that the greater departure of the residuals from Normality for the regression of OI compared with $\log(OI)$ is clearer than in Figure 5.

5.3 Envelope plots

The idea is to provide an envelope of points which will give some guidance to the analyst trying to decide whether observed departures from the expected line represent anything more than random fluctuation. While the envelopes are motivated by ideas from hypothesis testing, the aim is more informal and exploratory. They were probably first used, in the context of spatial statistics, by Ripley (1977) but their application to regression appeared in Atkinson (1981) and received further exposure in Cook and Weisberg (1982, p.56), Atkinson (1982) and Atkinson (1985). They were quite fashionable in the 1980s but are something of a rarity these days.

Envelope plots can be used for residuals that are scale-free, so apply to scaled, standardized and deletion residuals but not the ordinary residuals $\hat{\epsilon}_i$. The idea is to augment the actual residual with residuals from 19 simulated regression. At each expected normal order statistic or half-normal score, the actual residual, together with the maximum and minimum residual at that abscissa from the 19 simulations, are plotted. Each simulation is conducted by regressing a sample of n standard Normal variables on the covariates used in the regression. At first this may seem odd but, as the residuals are independent of scale and location, the fact that $\boldsymbol{y}$ and the standard normal variables have different scales and locations is immaterial. These simulations are, in effect, a device for producing residuals that have the same scale,

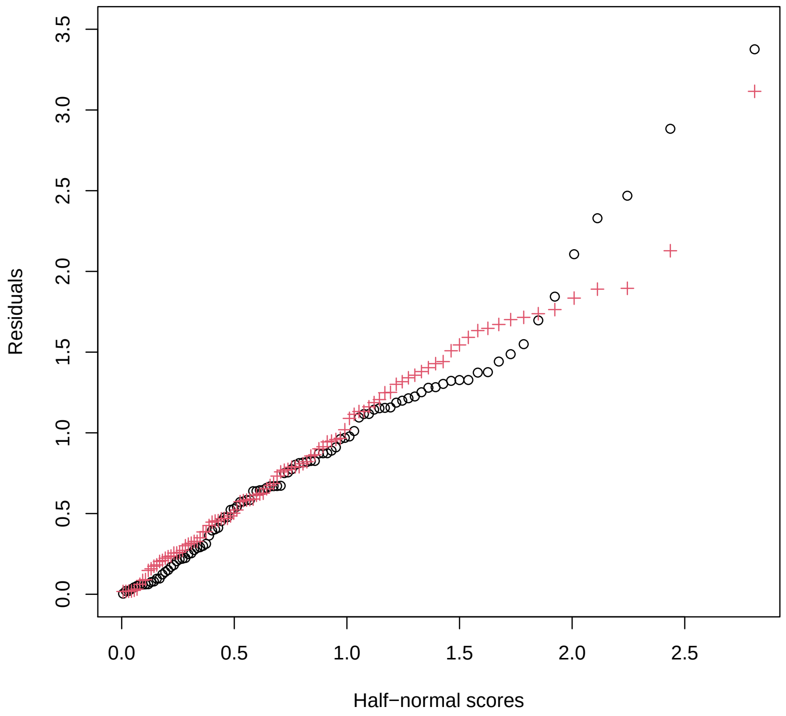
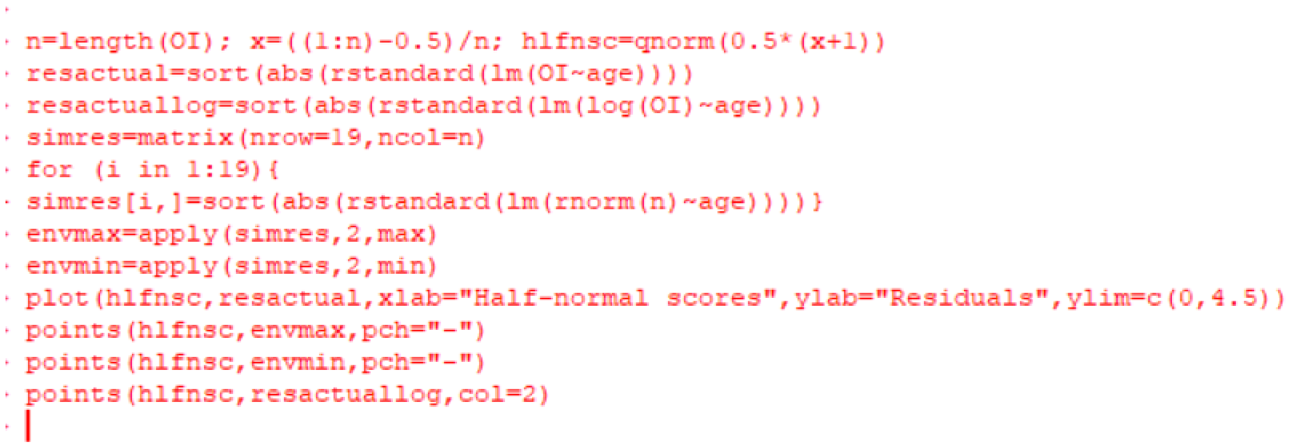


Figure 8: Half-normal plots of standardised residuals for the regression of OI on age (black circles) and $\log(OI)$ on age (red crosses)

location and dispersion as the actual residuals, provided the model is correct.

The R code for producing an envelope plot for the regression of OI on age is shown below. As the simulated dependent variables are simply $N(0, 1)$ variables, the envelope can be used for this model and for the regression of $\log OI$ on age, so this is included in the final line of the code.



```
n=length(OI); x=((1:n)-0.5)/n; hlfnsc=qnorm(0.5*(x+1))
resactual=sort(abs(rstandard(lm(OI~age))))
resactuallog=sort(abs(rstandard(lm(log(OI)~age))))
simres=matrix(nrow=19,ncol=n)
for (i in 1:19){
simres[i,]=sort(abs(rstandard(lm(rnorm(n)~age))))}
envmax=apply(simres,2,max)
envmin=apply(simres,2,min)
plot(hlfnsc,resactual,xlab="Half-normal scores",ylab="Residuals",ylim=c(0,4.5))
points(hlfnsc,envmax,pch="-")
points(hlfnsc,envmin,pch="-")
points(hlfnsc,resactuallog,col=2)
```

The resulting plot is shown in Figure 9, with the red points consistently within the envelope, whereas the black points are consistently below the envelope for half-normal scores between 1.5 and 2.

6 General remarks

Leaving aside specialised applications, such as the construction of centile charts, most analysts will use the foregoing material to check the data broadly conform to the assumptions of Normality in the model being fitted. In doing so, discrepancies will often take one of two forms - local or global departures: global departure are where the form of the model itself is under question, whereas local departures are where only a few points do not seem to fit the model. With global departures it is important not to be too demanding. Some issues, such as skewness, can be addressed with a log transformation and this can lead to better and more interpretable models. Remedial measures may be less obvious with other forms of departure but, in many cases, such departures may be less troublesome: arguments based on the Central Limit Theorem, and the inherent robustness of many techniques may mean that such forms of departure might safely be overlooked.

Caution is especially important in the use of more formal methods, such as the SW and SF tests. While the significant result obtained when OI was regressed on age, and the non-signifcant result when $\log OI$ was used instead, provided reassurance regarding the decision to analyse $\log OI$, graph-

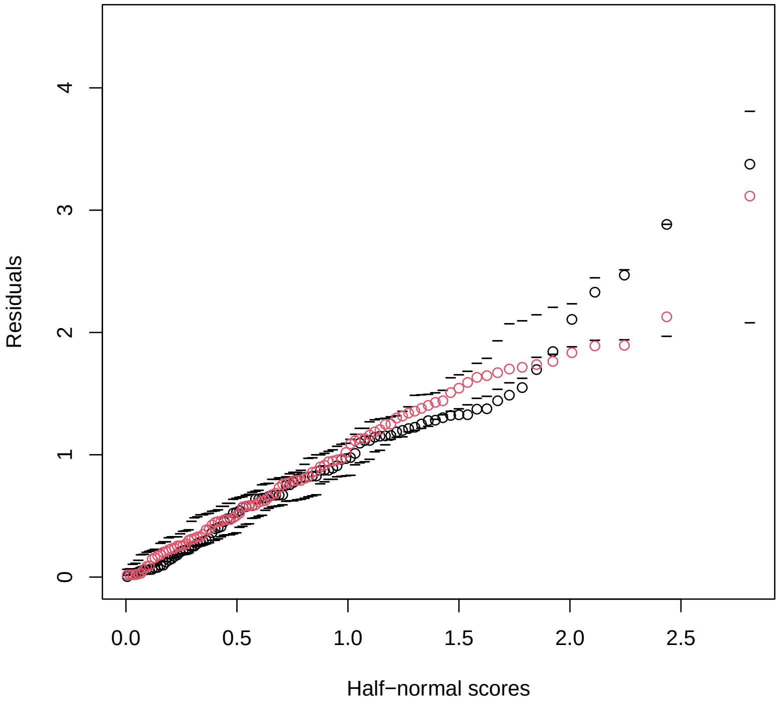


Figure 9: Half-normal plots of standardised residuals for the regression of OI on age (black circles) and $\log(OI)$ on age (red circles) with envelope from 19 simulated regressions

ical methods would have been sufficient to suggest the need for transformation. In small samples both SW and SF have low power, so non-significant results will be uninformative, while in large samples modest and unimportant departures from Normality may give a significant result.

Quite often the main value in the application of diagnostic plots, including NPPs and HNPs, is in the identification of outliers or other unusual points. Of course, once identified, aberrant data need investigating and careful handling. It may sometimes be appropriate to omit such points from a model if their inclusion has a material effect on the values and interpretation of key parameters, while ensuring that the omitted points are reported separately and the reasons for exclusion from the model documented.

This can be especially awkward when analysing data from clinical trials, where the doctrine of analysis which has evolved over the last twenty or so years regards omission of data points as a potent source of bias. While broadly understandable, current dogma sometimes seems less troubled by the possible distortion of treatment estimates caused by the inclusion of plainly aberrant data. A general prescription for managing this issue is probably not possible, and certainly unwise. When the issues are considered in the context of a particular study, the scientifically appropriate approach, or approaches, may be readily apparent. That these approaches may be difficult to reconcile with current practice is not a reason to eschew diagnostic techniques which have evolved over decades.

References

A. C. Atkinson. Two graphical displays for outlying and influential observations in regression. *Biometrika*, 68:13–20, 1981.

A. C. Atkinson. Regression diagnostics, transformations and constructed variables (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 44:1–36, 1982.

A C Atkinson. *Plots, Transformations, and Regression.* Oxford University Press, Oxford, 1985.

A Azzalini and A Capitano. *The Skew-Normal and Related Families.* Cambridge University Press, Cambridge, 2014.

Gunnar Blom. *Statistical estimates and transformed Beta-variables.* Wiley, New York, 1958.

R Dennis Cook and Sandford Weisberg. *Residuals and Influence in Regression.* Chapman and Hall, London, 1982.

Cuthbert Daniel. Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1:311–341, 1959.

J F Gentleman and M B Wilk. Detecting outliers in a two-way table: I. statistical behavior of residuals. *Technometrics*, 17:1–14, 1975.

M. Hills. On looking at large correlation matrices. *Biometrika*, 56:249–253, 1969.

N L Johnson, S Kotz, and N Balakrishnnan. *Continuous Univariate Distributions*, volume 1. Wiley, New York, 1994.

W. Königer. Remark AS R47: A remark on AS 177. expected normal order statistics (exact and approximate). *Applied Statistics*, 32:223–224, 1983.

B. D. Ripley. Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:172–212, 1977.

J P Royston. An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31(2):115–124, 1982a.

J P Royston. Algorithm AS 177. expected normal order statistics (exact and approximate). *Applied Statistics*, 31:161–165, 1982b.

S S Shapiro and R S Francia. An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67:215–216, 1972.

S S Shapiro and M B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1965.

Henry C Thode, Jr. *Testing for Normality.* Marcel Dekker, New York, 2002.

S Weisberg and C Bingham. An approximate analysis of variance test for non-normality suitable for machine calculation. *Technometrics*, 17:133–134, 1975.

A Joint distribution of two order statistics

The joint density of $X_{(r)}$ and $X_{(s)}$, $s > r$, $f_{rs}(x, y)$, can be derived constructively. For $x < y$, the probability $X_{(r)} \in (x, x + dx)$ and $X_{(s)} \in (y, y + dy)$ is found as the probability that this event occurs and it is X_1 that is $X_{(r)}$ and X_2 that is $X_{(s)}$ and then by multiplying by $n(n-1)$, as any ordered pair from $X_1, \ldots, X_n$ could have taken the role of X_1 and X_2.

The probability that $X_1 \in (x, x+dx)$ and $X_2 \in (y, y+dy)$ and that these are, respectively, the rth and sth order statistics, requires that of the remaining $n-2$ elements of the sample $r-1$ are less than x, $s-r-1$ must have values between x and y and the remaining $n-s$ are greater than y. The probabilities of an element of the sample falling in these intervals is $F(x), F(y) - F(x)$ and $1 - F(y)$, respectively. As the elements of the unordered sample are independent, this gives (omitting the dx and dy),

$$\begin{aligned} f_{rs}(x,y) &= n(n-1) \times f(x)f(y) \\ &\quad \times \frac{(n-2)!}{(r-1)!(s-r-1)!(n-s)!} F(x)^{r-1}[F(y)-F(x)]^{s-r-1}[1-F(y)]^{n-s}. \end{aligned}$$

The second line comes from the classification of the $n-2$ values into the three intervals, namely less than x, between x & y and greater than y, which is a multinomial probability (cf.1.4). Incorporating the $n(n-1)$ into the numerator of the multinomial coefficient gives the expression in (5). Clearly $f_{rs}(x,y) = 0$ if $x > y$.

B Some useful properties of $\boldsymbol{V}$ and $\boldsymbol{m}$

Start with the observation that if Z has a distribution symmetric about 0, then $-Z$ has the same distribution as Z. If $\boldsymbol{Z}$ denotes the vector of an *ordered* random sample from such a distribution, then the ordering means that $-\boldsymbol{Z}$ does not have the same distribution as $\boldsymbol{Z}$, but if we reversed the order of the elements of $-\boldsymbol{Z}$, then this would have the same distribution as $\boldsymbol{Z}$ - e.g. $Z_{(1)}$ has the same distribution as $-Z_{(n)}$. This can be expressed more precisely using the matrix $\boldsymbol{R}$ which reverses the order of a vector, e.g. for the 5×5 case:

$$\boldsymbol{R}_5 = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Note that $\boldsymbol{R}^T = \boldsymbol{R}$ and $\boldsymbol{R}\boldsymbol{R} = \boldsymbol{R}^2 = \boldsymbol{I}$, i.e. $\boldsymbol{R}$ is self-inverse. Then the above argument can be expressed as $\boldsymbol{Z}$ having the same distribution as $-\boldsymbol{R}\boldsymbol{Z}$. Consequently

$$\mathsf{E}[\boldsymbol{Z}] = \mathsf{E}[-\boldsymbol{R}\boldsymbol{Z}] \Rightarrow \boldsymbol{m} = -\boldsymbol{R}\boldsymbol{m}$$

and

$$\mathsf{var}[\boldsymbol{Z}] = \mathsf{var}[-\boldsymbol{RZ}] \Rightarrow \boldsymbol{V} = \boldsymbol{RVR}.$$

This shows that the mean is skew-symmetric, i.e. the vector reversed is the negative of the original vector, and the dispersion matrix is unchanged by reversing the order of the columns and the rows. It follows, e.g., that $\mathbf{1}^T\boldsymbol{m} = 0$ and $\boldsymbol{V}^{-1} = \boldsymbol{RV}^{-1}\boldsymbol{R}$. An important result for the SW test is that $\boldsymbol{a} = \boldsymbol{V}^{-1}\boldsymbol{m}$ is skew-symmetric. This follows by noting that $\boldsymbol{Ra} = \boldsymbol{RV}^{-1}\boldsymbol{RRm} = -\boldsymbol{V}^{-1}\boldsymbol{m}$.

Another result which helps to simplify and understand the calculations underpinning the SW test is that for a standard Normal distribution is that $\boldsymbol{V}\mathbf{1} = \mathbf{1}$. For this result, Normality is needed because symmetry alone is insufficient - e.g. it does not apply to samples from a t-distribution. The argument needs a little development and uses Basu's Theorem (cf. 1.6) and the independence of the sample mean and variance (cf. 1.2).

Suppose $X_1, \ldots, X_n$ are i.i.d. variables from a Normal distribution with mean μ and variance σ^2, Then the sample mean, $\overline{X}$ and sample variance s^2 are complete and sufficient for μ, σ^2. Also, the distribution of the statistic

$$U = \left(\frac{X_{(1)} - \overline{X}}{s}, \ldots, \frac{X_{(n)} - \overline{X}}{s}\right),$$

does not depend on μ or σ^2. This is because U is scale and location invariant. Therefore, by Basu's Theorem U and $(\overline{X}, s^2)$ are independent. As $\overline{X}$ and s^2 are independent, $\overline{X}$ is independent of sU, i.e. $\overline{X}$ is independent of $X_{(r)} - \overline{X}$, for any $r = 1, \ldots, n$. Consequently, $\mathsf{cov}[\overline{X}, (X_{(r)} - \overline{X})] = 0$. Applying this to a standard Normal sample $(Z_1, \ldots, Z_n)$ we get $\mathsf{cov}[\overline{Z}, Z_{(r)}] = \mathsf{var}[\overline{Z}] = n^{-1}$. It follows that:

$$\begin{aligned}\mathsf{E}[(\sum_i Z_{(i)})(Z_{(r)} - m_r)] &= \sum_i \mathsf{E}[Z_{(i)}(Z_{(r)} - m_r)] \\ &= \sum_i \mathsf{cov}[Z_{(i)}, Z_{(r)}] = \mathsf{cov}[n\overline{Z}, Z_{(r)}] = 1,\end{aligned}$$

where we have used result 1.5. As this holds for all $r = 1, \ldots, n$, we have $\boldsymbol{V}\mathbf{1} = \mathbf{1}$. From this it follows immediately that $\boldsymbol{V}^{-1}\mathbf{1} = \mathbf{1}$.