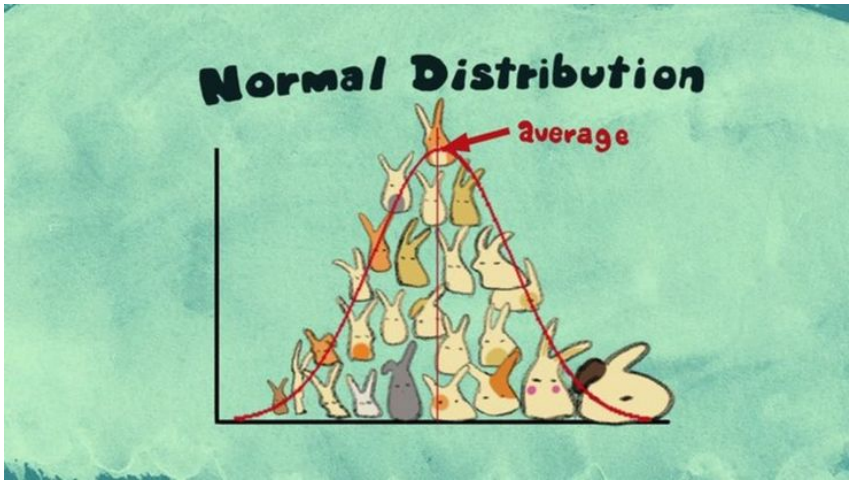
Normal Probability Plots & Order Statistics

J N S Matthews

Biostatistics Research Group, Newcastle University



Assessing Normality of data

Several ways to proceed:

1. Could plot a histogram - needs number of bins to be specified
2. Third and fourth moments of a Normal distribution are known:
   - Skewness is defined as:

     $$\frac{\mathsf{E}[(x-\mu)^3]}{\sigma^3}$$

     and vanishes for a Normal distribution
   - Kurtosis is defined as:

     $$\kappa = \frac{\mathsf{E}[(x-\mu)^3]}{\sigma^4}$$

     and is 3 for a Normal distribution ($\kappa - 3$ is the *excess* Kurtosis). Measures heaviness of tails of distribution. *t*-distribution has $\kappa > 3$ and is *leptokurtic* (as opposed to *platykurtic*)

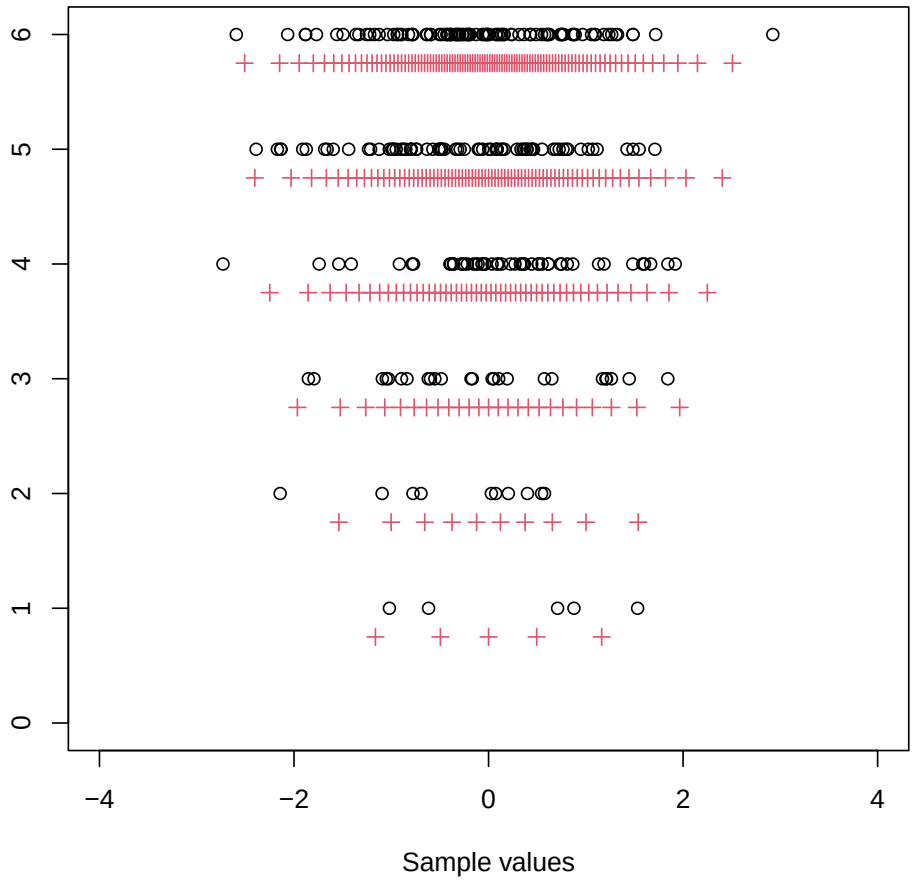
Normal Probability Plots

An alternative graphical method, requiring no specification of bins etc. is the *Normal Probability Plot* (NPP)

- Method is based on *ordering* the data
- Suppose $Z_1, \ldots, Z_n$ is a sample of n independent realisations from a N(0,1) distribution.
- Once ordered, sample is denoted by $Z_{(1)} < Z_{(2)} < \ldots < Z_{(n)}$ - the *order statistics*.
- Can compute the *expected order statistics*, $\mathsf{E}[Z_{(i)}]$ for $i = 1, \ldots, n$
- Realisations of $Z_{(i)}$ and values of $\mathsf{E}[Z_{(i)}]$ are shown on next slide for various sample sizes

Example order statistics

Samples of sizes 5,10,25,50,75,100 - expected values in red

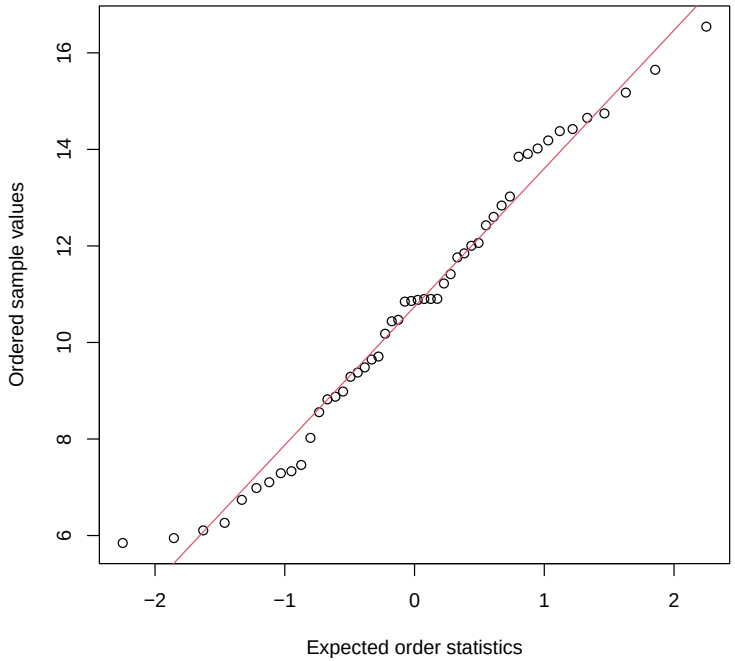


Practical Use

1. Confronted with some real data, $Y_1, \ldots, Y_n$, how can this be used to assess Normality?
2. Note that $Y_i = \mu + \sigma Z_i$ translates to $Y_{(i)} = \mu + \sigma Z_{(i)}$ because $\sigma > 0$.
3. So plot of $Y_{(i)}$ versus $\mathrm{E}[Z_{(i)}]$ should give a line that is approximately straight and
4. a regression of $Y_{(i)}$ on $\mathrm{E}[Z_{(i)}]$ will have slope σ and intercept μ

Example of NPP

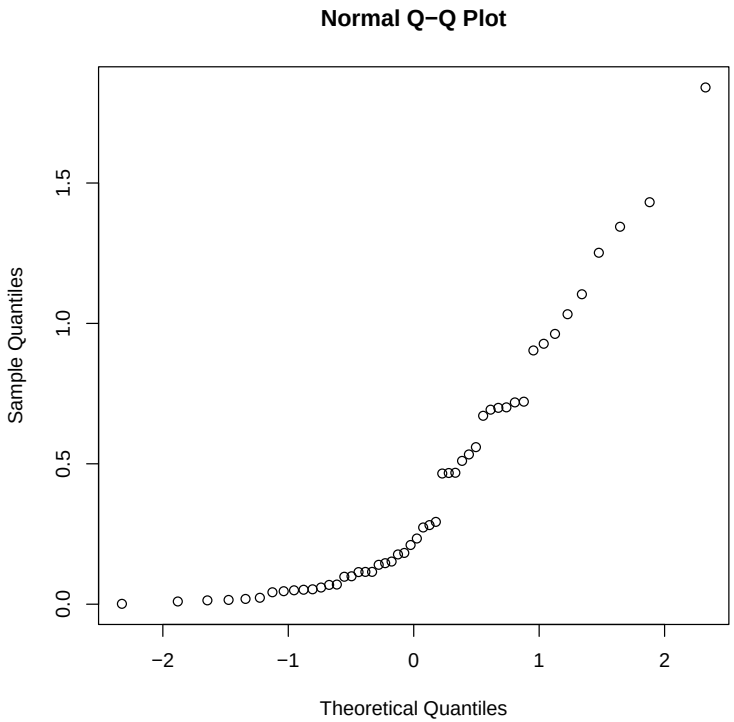
Plot is for a random sample, size 50, from N$(10, 3^2)$.

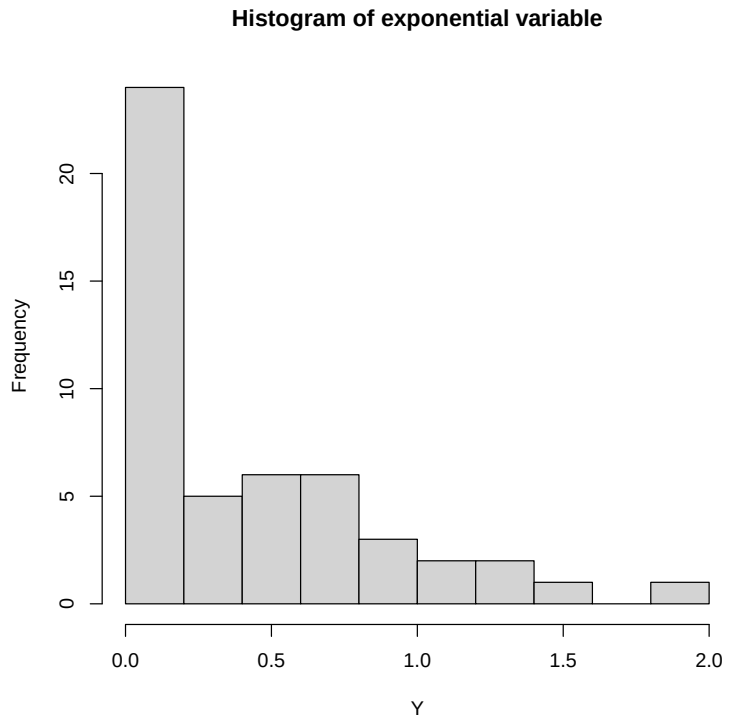


Simple regression line in red - intercept 10.74, slope 2.87

How well do NPPs pick up non-Normality?

NPP & histogram of an exponential variable mean $\frac{1}{2}$

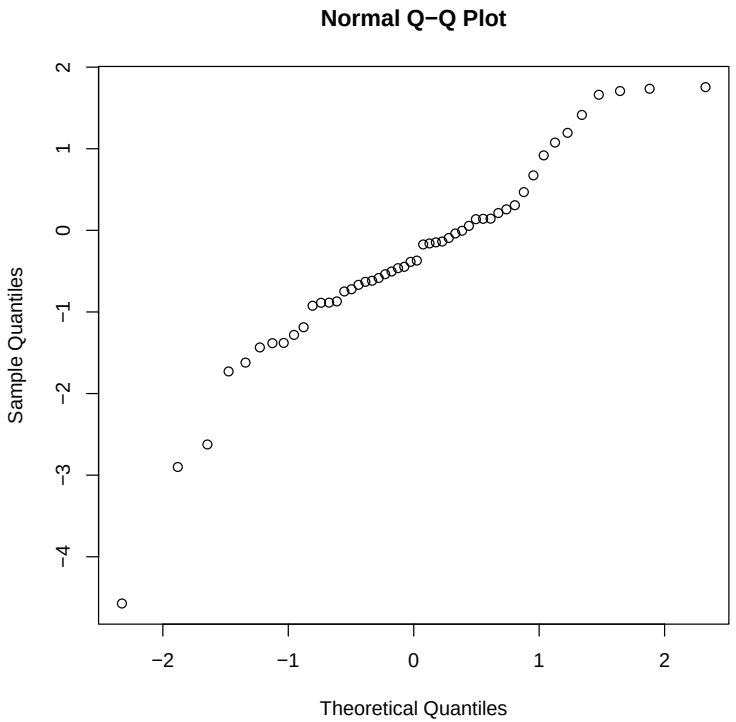


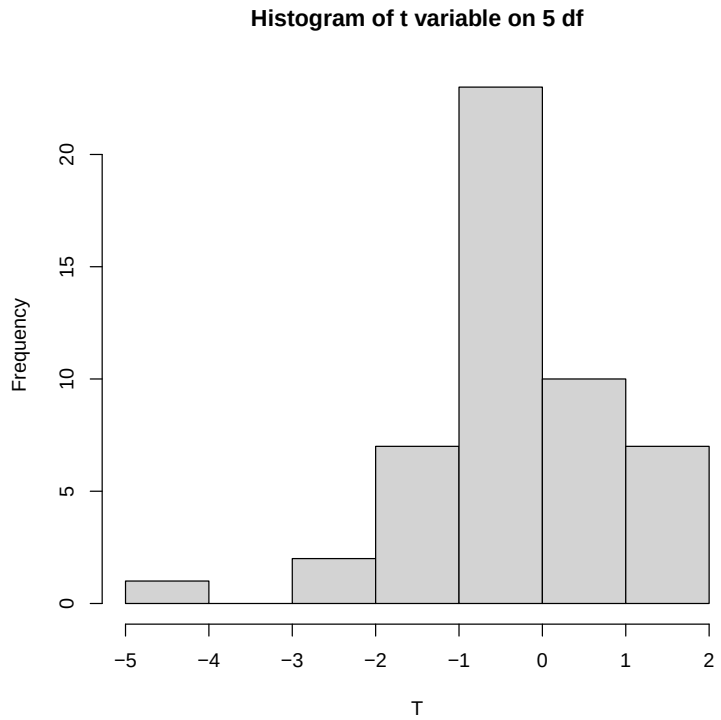


Histogram and NPP seem equally good - but this is an easy case

Symmetric, non-Normal case

NPP & histogram of a t-variable with 5 df, $\kappa = 9$





Histogram and NPP seem to struggle equally - symmetric & non-Normal not easy to detect - perhaps tails of NPP more telling

Order statistics - general properties

To plot a NPP we need $\mathrm{E}[Z_{(i)}]$ - how is this computed?

1. The X_i have distribution and density $F(\cdot)$ and $f(\cdot)$, respectively.
2. Distribution function of $X_{(r)}$ is $F_r(\cdot)$.
3. The event $\{X_{(r)} < x\}$ occurs when at least r of the sample are less than x.
4. Occurs with probability $\sum_{j=r}^{n} a_j$, where

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

5. To get the density of $X_{(r)}$, $f_r(x)$, this must be differentiated.

Density of $X_{(r)}$

Now $\frac{da_j}{dx} = b_j - c_j$, where

$$b_j = j \binom{n}{j} F(x)^{j-1}[1 - F(x)]^{n-j} f(x)$$

$$c_j = (n-j) \binom{n}{j} F(x)^{j}[1 - F(x)]^{n-j-1} f(x)$$

and it turns out $b_{j+1} = c_j$ and $c_n = 0$, so the sum telescopes, giving

$$f_r(x) = r \binom{n}{r} F(x)^{r-1}[1 - F(x)]^{n-r} f(x)$$

Also, note there is an alternative form because

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

Order statistics, Uniform distribution

1. General form of $f_r(\cdot)$ not that useful
2. but it is for the Uniform distribution with $F(x) = x$ and $f(x) = 1$
3. Substituting in general result shows $U_{(r)}$ has a Beta distribution with parameters r and $n - r + 1$,
4. which has mean

$$\mathsf{E}[U_{(r)}] = \frac{r}{n+1}$$

Order statistics, Normal distribution

1. Expression for $f_r(x)$ with $F = \Phi$ and $f = \phi$ does not provide tractable moments. So need approximations.
2. Note that $\Phi(Z_r) = U_r$, and as Φ monotone increasing, $\Phi(Z_{(r)}) = U_{(r)}$.
3. Initial attempt might be to use

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

4. Not entirely without justification - it is a first order Taylor expansion.

Better approximations

1. Not a brilliant approximation. Rather than use higher order Taylor series, Blom (1958) considered

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

2. $\alpha = 0$ is foregoing case, $\alpha = \frac{1}{2}$ is often seen, but best compromise that is independent of n is $\alpha = \frac{3}{8}$
3. R routines are available to get very accurate results using numerical integration.

Comparison of approximations

Examples for $n = 50$

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
⋮				
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

Dependence of order statistics

1. Even if the X_i are independent, the $X_{(i)}$ are not.
2. $\Pr(X_1 < x \mid X_2 < x) = \Pr(X_1 < x)$
3. but clearly
$$\Pr(X_{(1)} < x \mid X_{(2)} < x) = 1$$
4. So $\text{cov}(X_{(r)}, X_{(s)})$ will be non-zero and that has some implications later. The joint density of these variables, $f_{rs}(x, y)$ can be found using an extension of the argument used for $f_r(x)$ - see the notes.

Tests for Normality

1. If we want to know whether data are Normal, then we could test the null hypothesis that the data are Normal.
2. Indeed we can. In fact books have been written on it (e.g. Thode, 2002)
3. How some methods work is fairly obvious: e.g. testing skewness or kurtosis or using Kolomogorov-Smirnov.
4. Several variants of a method that turns out to be a good omnibus test (i.e. good against a range of alternatives), associated with the names of Shapiro-Wilk (SW), Shapiro-Francia (SF), Weisberg-Bingham are, perhaps, in need of more explanation.

Should we test for Normality?

- The following slides will explain the SW and SF test. But should we bother?
- Most tests have poor power in moderate sized samples. Tests focussed on a given aspect, such as skewness, may have better power against that aspect than, say SW or SF, but will miss other aspects entirely.
- Poor power means that failure to discredit Normality may mean little.
- But for many purposes, modest (or even larger?) departures from Normality will not be too troublesome. Most methods are very forgiving is this respect.
- Some applications rely more heavily on the Normality assumption, such as determination of reference ranges or centile charts.

Shapiro-Wilk Test

1. The SW test based on the following observations
2. If the data are Normal, then the slope of the NPP is an estimator of σ.
3. Regardless of the Normality of the data, the usual sample SD also estimates σ.
4. SW works by comparing these estimates (although the actual test statistic uses a scaling that isn't entirely transparent)

Estimate of σ

1. Define $m_i = \mathrm{E}[Z_{(i)}]$ and $\boldsymbol{m}$ as vector of m_i, and $\mathbf{1}$ as vector of ones. Then, if $\boldsymbol{y}$ are the *ordered* data

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

2. Note use of generalized least squares with weighting matrix the inverse of $\boldsymbol{V}$, where $(\boldsymbol{V})_{rs} = \mathrm{cov}[X_{(r)}, X_{(s)}]$
3. Numerical methods based on $f_{rs}(x, y)$ needed to obtain $\boldsymbol{V}$
4. Some surprising simplifications are possible

A methodological excursion - Basu's Theorem

Theorem

Basu's Theorem: *suppose data are from a family with density* $f(\cdot \mid \boldsymbol{\theta})$. *If* T *is complete and sufficient for* $\boldsymbol{\theta}$ *and* V *has a distribution that does not depend on* $\boldsymbol{\theta}$, *then* T *and* V *are independent.*

And also:

Theorem

If data are Normally distributed, then the sample mean, $\overline{X}$, *and sample variance,* s^2, *are independent.*

A curious result

1. For Normal data, $(\overline{X}, s^2)$ is complete and sufficient for (μ, σ^2).
2. $\left(\frac{X_{(1)}-\overline{X}}{s}, \ldots, \frac{X_{(n)}-\overline{X}}{s} \right)$ clearly does not depend on (μ, σ)
3. So $(X_{(1)} - \overline{X}, \ldots, X_{(n)} - \overline{X})$ is independent of $\overline{X}$
4. Consequently for standard Normal variables, $\mathrm{cov}[Z_{(r)}, \overline{Z}] = \mathrm{var}[\overline{Z}]$, for $r = 1, \ldots, n$
5. Using $n\overline{Z} = \sum Z_{(i)}$ this leads to $\boldsymbol{V}\mathbf{1} = \mathbf{1}$ and hence $\boldsymbol{V}^{-1}\mathbf{1} = \mathbf{1}$

Back to SW

1. By symmetry $\boldsymbol{m}^T\mathbf{1} = 0$, which with $\boldsymbol{V}^{-1}\mathbf{1} = \mathbf{1}$ gives

$$\hat{\mu} = \overline{\boldsymbol{y}}$$

$$\hat{\sigma} = \frac{\boldsymbol{m}^T\boldsymbol{V}^{-1}\boldsymbol{y}}{\boldsymbol{m}^T\boldsymbol{V}^{-1}\boldsymbol{m}}$$

2. Shapiro & Wilk took $\boldsymbol{a}$ to be a unit vector proportional to $\boldsymbol{V}^{-1}\boldsymbol{m}$ and defined the SW statistic as:

$$W = \frac{(\boldsymbol{a}^T\boldsymbol{y})^2}{(n-1)s^2} = \frac{(\boldsymbol{m}^T\boldsymbol{V}^{-1}\boldsymbol{y})^2}{(n-1)s^2\boldsymbol{m}^T\boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{m}}$$

Some properties of W

1. An application of Cauchy-Schwartz shows that $0 \leq W \leq 1$
2. In the notes $\mathrm{E}[W]$ is derived: for $n = 50$, $\mathrm{E}[W] = 0.967$.
3. In the null case W is a negatively skewed variable, with values further from one corresponding to non-Normal data
4. p-values can be found by simulation these days, although Royston (1982) applied a Box-Cox transformation to $1 - W$ in order to derive p-values. Can use `shapiro.test` in R .

Shapiro-Francia (SF) test

The SW statistic is awkward to calculate because it requires the evaluation of $\boldsymbol{V}$.

1. This is required because it is (essentially) the dispersion matrix of the ordered data, so we use GLS
2. If we ignored this feature and used OLS the estimators would still be consistent.
3. This is the approach of the Shapiro-Francia statistic, W_f
4. Same expression but with $\boldsymbol{V}$ replaced with the identity
5. That is

$$W_f = \frac{(\boldsymbol{b}^T \boldsymbol{y})^2}{(n-1)s^2}$$

where $\boldsymbol{b}$ is a unit vector proportional to $\boldsymbol{m}$.

SF test and further simplification

1. As with W, $0 \leq W_f \leq 1$, with small W_f discrediting null hypothesis
2. Indeed, if $\boldsymbol{y} = \bar{y} + s\boldsymbol{m}$, then $W_f = 1$.
3. Routine for SF is in R library `DescTools`.
4. SF is still a little awkward as $\boldsymbol{m}$ is still required. Weisberg and Bingham used $\Phi^{-1}[(i - \frac{3}{8})/(n + \frac{1}{4})]$ in place of m_i.
5. SW, SF and Weisberg-Bingham have very similar properties

Application to residuals

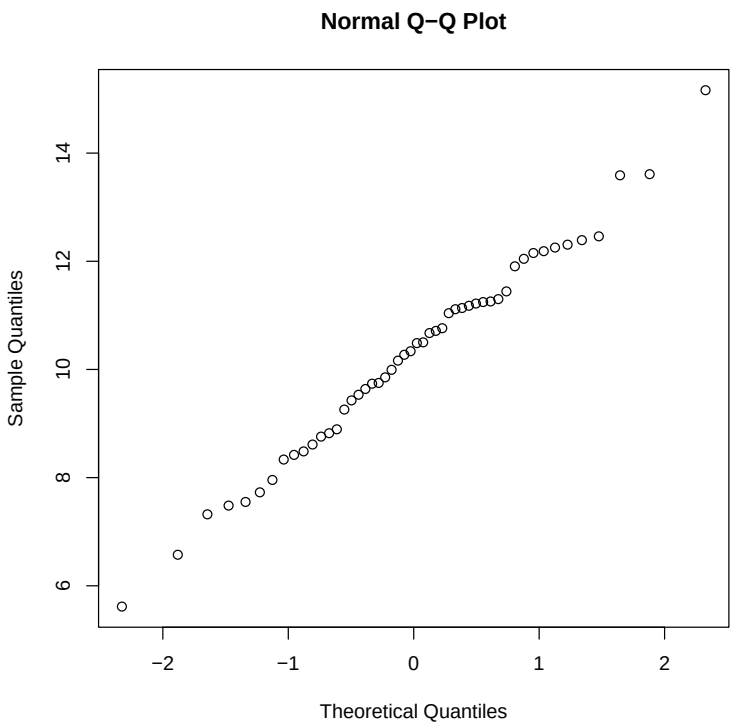
- Could apply any of these tests to estimated residuals
- However, estimated residuals are not independent
- Complexity of SW test arises from taking account of dependence of observations *induced by the ordering*
- Estimated residuals are dependent even before ordering.
- More detail in notes - but probably no case at all for using W rather than simpler versions in this application.
- Also note that $\hat{\epsilon} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon}$, i.e. *estimated* residuals are linear combinations of residuals
- Can exaggerate their Normality - idea of *supernormality*.

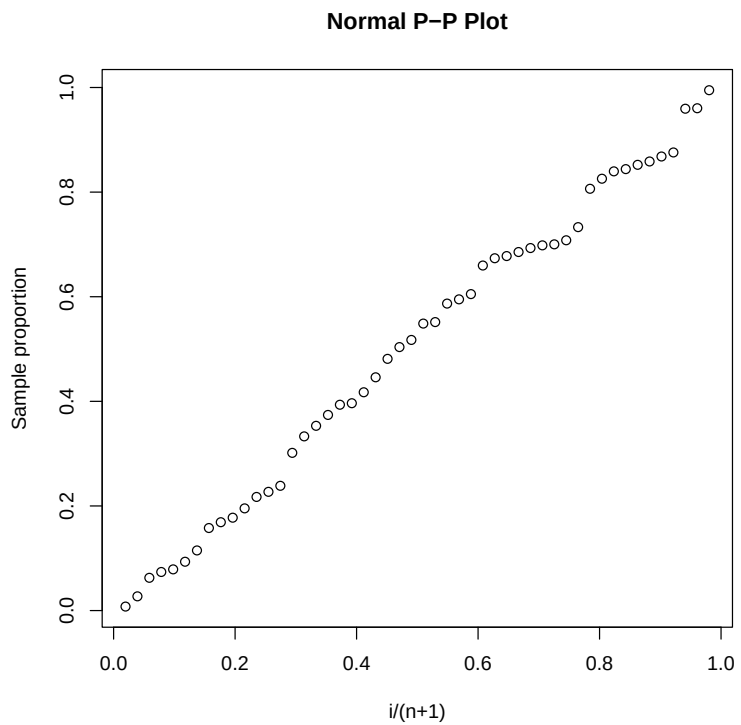
Q-Q and P-P plots

1. NPPs are examples of Q-Q plots
2. Q-Q plots plot quantiles of data against those of putative distribution - ranges are those of the support of the data.
3. I.e. $Y_{(i)}$ vs $F^{-1}(\mathsf{E}[U_{(i)}]) = F^{-1}(i/(n+1))$.
4. P-P plots transform to a 0-1 scale, i.e. $F(Y_{(i)})$ vs $i/(n+1)$.
5. If null is true, graph aligns along $y = x$
6. Cannot estimate parameters from graph - for Normal P-P plot, ordinate, $\Phi((Y_{(i)} - \hat{\mu})/\hat{\sigma})$ requires parameter estimates.
7. Probably easier to spot outliers in a Q-Q plot

Q-Q & P-P plot examples

Q-Q and P-P plots for sample ($n = 50$) generated by `rnorm`





Half-Normal plots

1. Half-Normal plots (HNPs) are Q-Q plots based on the Half-Normal distribution
2. The Half-Normal distribution is the distribution of $|Y|$, where $Y \sim N(0, \sigma^2)$. A standard Half-Normal is the distribution of $|Z|$.
3. The distribution function of $|Z|$ is $2\Phi(x) - 1, x > 0$.
4. Useful to assess variables that are expected to be zero-mean Normal variables, such as residuals

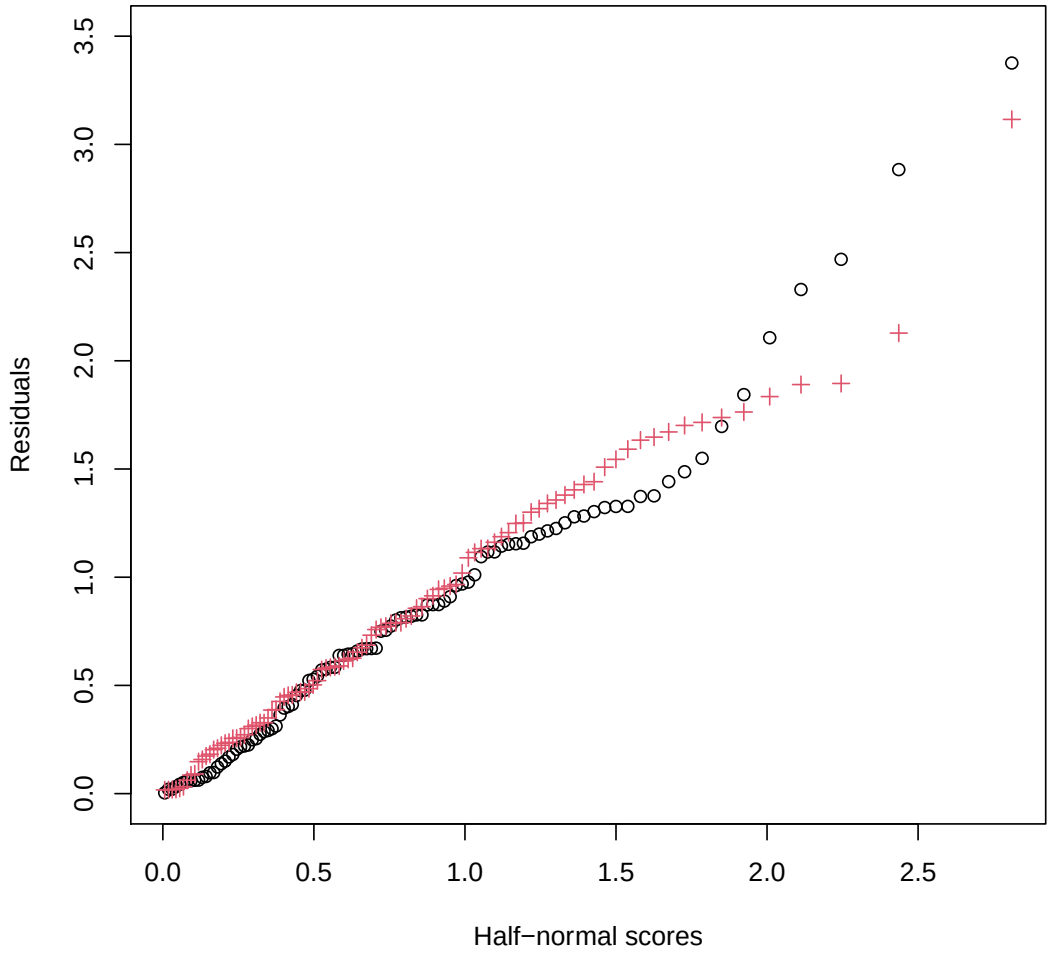
Uses and using HNPs

1. First used for assessing effects in high-order factorial designs
2. Can be used to help identify non-zero correlations in a large correlation matrix.
3. Plot ordered absolute values against $\Phi^{-1}(\frac{1}{2}(x_i+1))$, where x_i could be

$$\frac{i}{n+1} \quad \text{or} \quad \frac{i-\frac{1}{2}}{n}$$

Example HNPs

The residuals from regressing either OI (black) or $\log OI$ (red) on age



Envelope plots

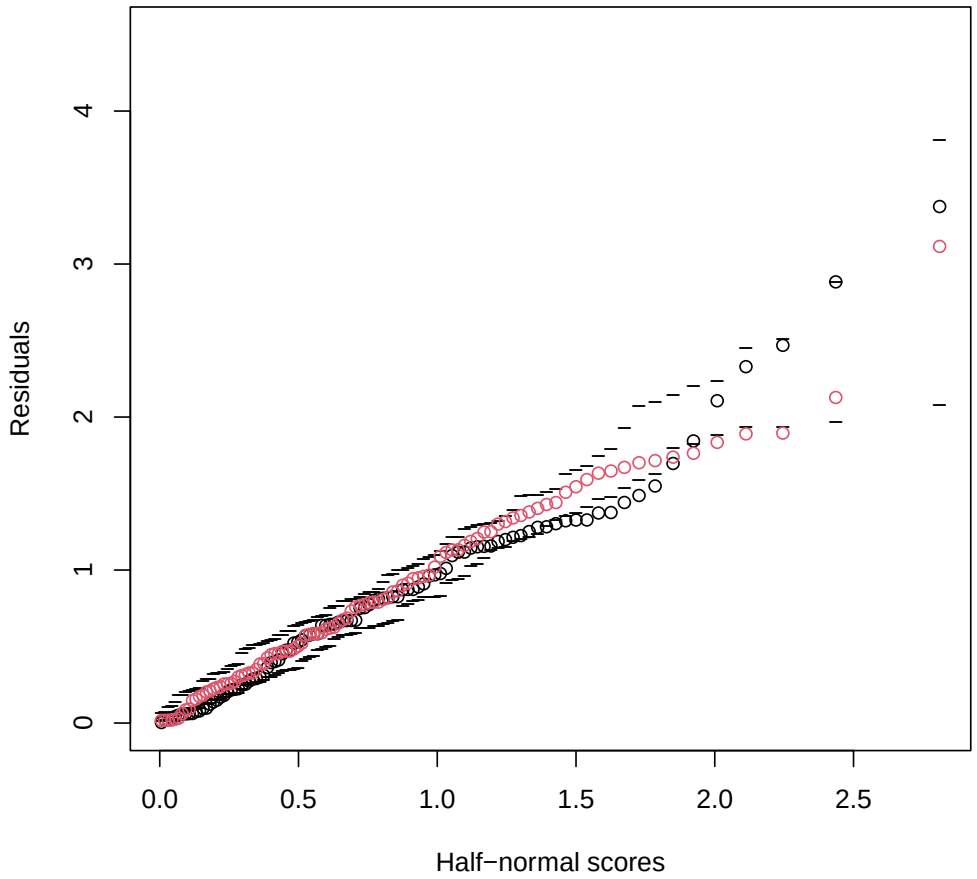
1. An abiding issue with the assessment of NPPs or HNPs is 'How straight is straight?'
2. Envelope plots, popular in '80s and '90s, provide some help
3. Idea is to generate a number of sets of residuals that are stochastically the same as the observed residuals (if model is correct)
4. Then plot their envelope as well as observed residuals [i.e. max and min of simulated residuals at each abscissa]

Generating envelope

1. Observation $Y_i \sim N((\boldsymbol{X}\boldsymbol{\beta})_i, \sigma^2)$ differs from $Z_i \sim N(0,1)$ only by scale and location.
2. *scale-free* residuals - i.e. scaled, standardized or deletion residuals will therefore be same if we regress either Y_i or Z_i on $\boldsymbol{X}$ [assuming model is correct]
3. So generate 19 sets of scale-free residuals using standard Normal variables for dependent variable and same covariates as in model. Use these to generate envelope
4. Really just a device to get a vector from $N(0, \boldsymbol{I} - \boldsymbol{P})$

Example of envelope plots

Envelope plots HNP of residuals from OI (black) and $\log OI$ (red) on age



Some remarks

1. Testing for Normality not easy [try running `hist(rnorm(10))` a few times in R]
2. Should always ask if it is necessary - depends on context
3. If not Normal what will you do? Skewness probably easiest violation to detect and, often, easy to remedy by using logs
4. Departures from Normality may be global but often local
5. Often NPPs etc. are best at spotting latter, i.e. outliers or points to be questioned.