Self-Calibrating Quantile–Quantile Plots

R. Wayne OLDFORD

Quantile–quantile plots, or qqplots, are an important visual tool for many applications but their interpretation requires some care and often more experience. This apparent subjectivity is unnecessary. By drawing on the computational and display facilities now widely available, qqplots are easily enriched to help with their interpretation. An overview of quantile functions and quantile–quantile plots is presented against the backdrop of their early historical development. Strengths and shortcomings of the traditional display are described. A new enhanced qqplot, the self-calibrating qqplot, is introduced and demonstrated on a variety of examples—both synthetic and real. Real examples include normal qqplots, log-normal plots, half-normal plots for factorial experiments, qqplots for $\bar{x}$ and s in process improvement applications, detection of multivariate outliers, and the comparison of empirical distributions. Self-calibration is had by visually incorporating sampling variation in the qqplot display in a variety of ways. The new qqplot is available through the function and R package `qqtest`.

KEY WORDS: Daniel plots; Half-normal plots; Multivariate outlier detection; Ogive; Visual hypothesis testing.; $\bar{x}$ and s charts

1. INTRODUCTION

> "I shall best explain my graphical method of expressing Distribution, which I like the more, the more I use it, and which I have latterly much developed, . . ."
>
> Francis Galton, *Natural Inheritance* (Galton 1889, p. 37)

Galton's graphical method is arguably the first quantile–quantile plot (qqplot) and has continued to be much developed over the nearly century and a half since its inception. It is of considerable value in practice and has become a staple method in all applied statistics courses.

The quantile curve can say a lot about the corresponding distribution, including detailed information such as percentiles, location, and scale, etc. The simple shape of its curve tells much about the distributional shape (e.g., symmetry/asymmetry, number of modes, tail weights, etc.). Quantile–quantile plots can directly compare one distribution to another (on any of these points) and so provide an informal visual test of whether observed values appear to follow some assumed distribution (e.g., the normal). Moreover, contrary evidence appearing in the plot reveals how the empirical distribution differs from that assumed.

As a visual test, problems of interpretation arise because no sense of sampling variability is contained in the plot itself. Instead, practitioners have come to rely on repeated experience to guide judgment. Much of this subjectivity is no longer necessary. Quantile–quantile plots can be substantively improved and made more easily and correctly interpretable by simply enhancing the displays with information on sampling variability. By contrasting the observed data against many that might have been generated by the assumed distribution, the enhanced plot builds in a measure of self-calibration.

In what follows, three separate enhancements for self-calibrated qqplots are described: the use of point-wise confidence envelopes, the overlaying of transparent exemplars, and the line-up plot that spatially separates exemplars. These enhancements are available as the `qqtest` package contributed to R (R Core Team 2014; Oldford 2015).

Section 2 provides a self-contained exposition on quantile functions and quantile–quantile plots to lay the groundwork needed for their understanding and interpretation. Historical episodes in the development of the plot from Galton (1875) to Hazen (1914) and Whipple (1916a, 1916b) are used to provide context and to illustrate their original practical applications and interpretation. The section ends with some discussion of the ambiguity inherent in using the traditional qqplot as a visual test.

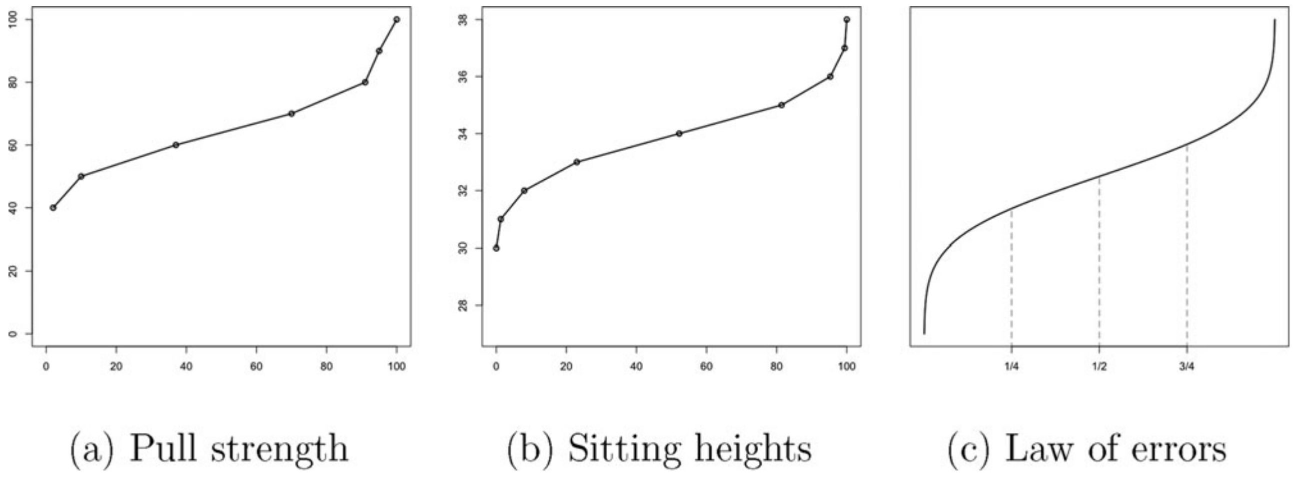
Section 3 introduces the self-calibrating quantile–quantile plot. Its construction is described and its performance under null and nonnull configurations is illustrated. Point-wise confidence envelopes are introduced as the default self-calibrating enhancement for `qqtest` that will be used throughout the article. Exemplars are also introduced and their use contrasted with that of the confidence envelopes—each provides a different quality of self-calibration.

The strength of the self-calibrated qqplot is best seen in its application in some meaningful context. To this end, in Section 4 it is applied to several and varied examples from the literature. These include the historical uses from Section 2 and

R. Wayne Oldford is Professor, Department of Statistics and Actuarial Science, 200 University Avenue West, University of Waterloo, Waterloo, ON N2L 3G1, Canada (E-mail: *rwoldford@uwaterloo.ca*). The author thanks the referees, associate editor, and editor for comments that resulted in a much improved article. The author also thanks his colleague Jock Mackay for helpful discussion on control charting and Christine Anderson-Cook who first encouraged him to write an article illustrating `qqtest`.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/tas*.

DOI: 10.1080/00031305.2015.1090338



(a) Pull strength (b) Sitting heights (c) Law of errors

Figure 1. Galton's "Ogive" (Galton 1875) or "Scheme of Distribution" (Galton 1889) or, in modern terms, "quantile function." Data shown are from measurements taken by Galton at the 1884 International Health Exhibition in London: (a) shows the pull strength in lbs. of 519 men aged 23–26 (source: Galton 1889, pp. 38, 199); (b) the sitting heights in inches of 775 women aged 23–50 (data source: Galton 1885a); (c) is the ogive given by the law of errors, viz., the quantile function of a normal distribution (as in Galton 1875).

other historically influential applications. The latter include the half-normal plots of Daniel (1959) for unreplicated factorial experiments and the exploratory multivariate outlier detection methods pioneered by Wilk and Gnanadesikan (1964). The remaining examples include a novel use of the self-calibrating qqplot in statistical process improvement and a visual approach to the classical two-sample problem.

In Section 5, the use of spatially separated exemplars is demonstrated. This is presented as an instance of the general visual significance testing method called a line-up plot (Buja et al., 2009). The article wraps up with a few concluding remarks in Section 6.

All data and methods presented are available in the qqtest R package. Readers are encouraged to download the package and try the examples themselves.

2. QUANTILE FUNCTIONS

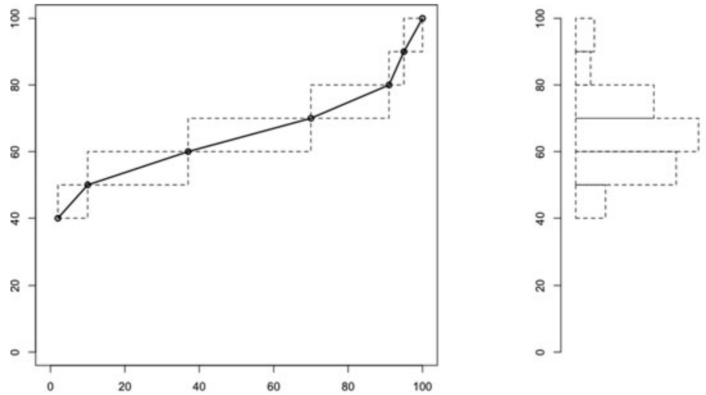
Galton's graphical innovation (Galton 1875, 1889) was to plot sample measurements, y_i say ($i = 1, \ldots, n$), on the vertical axis against their (smallest to largest) rank, r_i say, on the horizontal axis, drawing a smooth curve through the points to aid in interpretation. Figure 1(a) and 1(b) shows the plot for some of the data Galton gathered in 1884 at London's International Health Exhibition (Galton 1885b, 1885c, 1889)—in these figures Galton plotted only selected percentiles of the data to determine the curve. The curvilinear shape seen in Figure 1(a) and 1(b) and in idealized form in Figure 1(c) suggested the name "ogive" to Galton, after the architectural shape of an ogival arch (a contemporary account of which would be Smith 1884, pp. xvi, 129). Today, we know it as the quantile function—the sample or empirical quantile function in Figure 1(a) and 1(b) and the theoretical normal quantile function in Figure 1(c).

Galton showed that by simply plotting the data from smallest to largest against their percent position in that order, any number of statistical summaries could be had at once. From the horizontal axis, the value of *any* percentile is determined by its coordinate on the curve and conversely the percentile of any observation can be determined by its value on the horizontal axis. Robust measures of location are easily had—the median is the middle point on the curve or, if that is unavailable, the median can be estimated from a line drawn between any two available percentile values on the plot, preferably nearer the middle (Galton 1899). For a measure of variation, or scale, half the interquartile range is equivalent to the "probable error" and requires only two further points to be accurately measured (one corresponding to the first quartile, the other the third). Alternatively, the slope of the central linear part of the curve could serve as a measure of scale. Any of these are much simpler to determine than would be the sample mean and standard deviation. Moreover, as Galton (1889, p. 47) noted, the plot can often still "be constructed from observations that are barely exact enough . . . to be called measures." In fact, for measures of location and scale one need not even measure all points in the sample—it is enough to be able to order them and then measure only those at the three quartile positions and so plot only the middle part of the curve (e.g., the heights of men as in Galton 1875).

The curve can also say something about the shape of the data distribution (i.e., beyond its location and scale already summarized by the curve's height and slope). As Galton (1889) noted, whatever such information that can be had from a "frequency distribution" (or histogram) can also be had directly from the shape of this curve. In a figure similar to Figure 2, Galton (1889) showed the ogive for pull strength overlaid with boxes marking the vertical and horizontal difference between points and then the boxes moved to a common vertical axis (as in Figure 2(b)). The boxes are thus seen to determine the bins of the corresponding histogram—horizontal distances on the ogive provide the bin height of the histogram, vertical distances provide the bin width. Relatively flat plateaus in the quantile function indicate a concentration of points in the distribution, a mode.

2.1 Non-Gaussian Quantile Functions

The first row of Figure 3 shows the quantile function for the random variable $Y \sim F_Y(y)$ for various cumulative distribution functions $F_Y(y) = \Pr(Y \leq y)$. More precisely, the



(a) Pull strength ogive (b) Pull strength histogram

Figure 2. The quantile function (ogive) and the histogram have the same information content as regards the shape of the data distribution. Adapted from figs. 2 and 3 on page 38 of Galton (1889).

quantile function $Q_Y(p)$ is defined to be

$$Q_Y(p) := \inf\{y \in \mathbb{R} : p \leq F_Y(y)\}$$

for $p \in (0, 1)$ and the plots show the parametric curves $(p, Q_Y(p))$. In the case of continuous F_Y, this yields $Q_Y(p) = F_Y^{-1}(p)$ (i.e., swapping the horizontal and vertical axes of any plot in the first row of Figure 3 will yield a plot of the corresponding distribution function $p = F_Y(y)$). The first of these curves shows the familiar ogival shape of the normal or Gaussian distribution but the rest do not—“ogive” no longer has any descriptive power.

The second row of Figure 3 shows Galton’s plots of the ordered values of y_i against their ranks r_i for a sample of size $n = 25$ drawn from each of these distributions. As can be readily seen, for these particular samples at least, the plot of the sample points roughly follows the same shape as that of the quantile functions immediately above them. Were we to connect the dots of these points, the resulting sample curve could be considered an estimate of the theoretical quantile function and denoted by $\widehat{Q}_Y(p)$ to emphasize this point. The estimate $\widehat{Q}_Y(p)$ could provide evidence to support or refute possible model choices for $Q_Y(p)$.

What evidence of distributional shape do the plots in Figure 3 provide? As with Galton’s explication of Figure 2 we could imagine equal width bins formed along the vertical axis in each plot and each bin’s height determined from horizontal box widths (or in the second row by the count of the number of points in that bin). The same information however can be read directly from the quantile plot.

All plots of Figure 3 have been given common vertical scales to focus attention on what they have to say about distributional shape. To that end, the curves in the first row of Figure 3 are reliable guides that show the distinct signatures of different distributional shapes. In every case, relatively flat regions in the curve indicate a mode in the distribution—a single mode is indicated in the first three plots, two in the fourth, and none in the

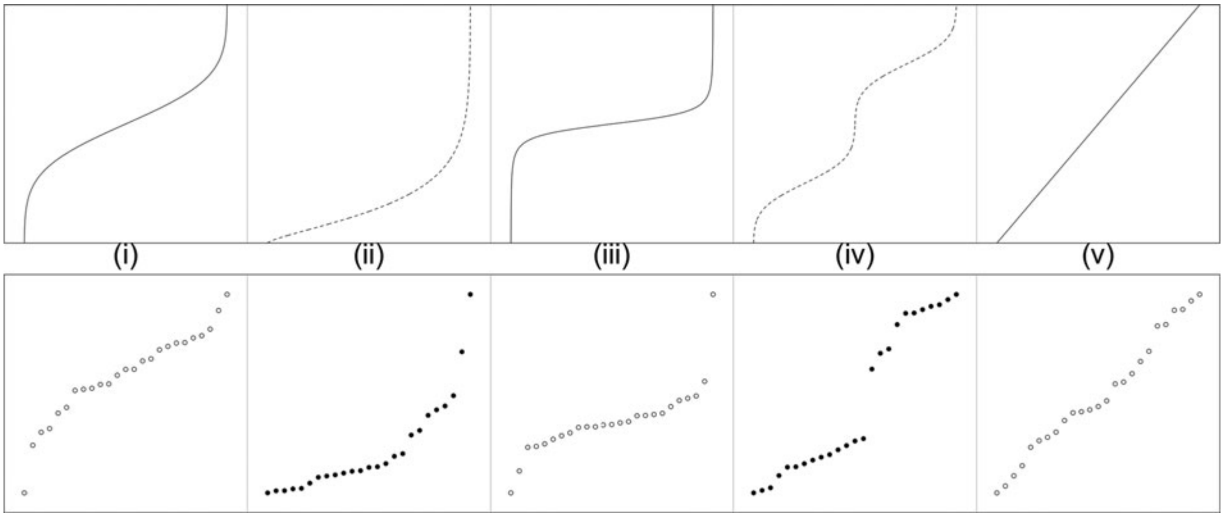


Figure 3. Quantile plots for various distributions: first row shows five theoretical quantile functions, second row shows samples of 25 ordered observations from the same distributions. (i) $Y \sim N(0, 1)$; (ii) $Y \sim \chi_3^2$; (iii) $Y \sim t_3$; (iv) $Y \sim \frac{1}{2}N(0, 1) + \frac{1}{2}N(6, 1)$; and (v) $Y \sim U(0, 1)$.

last. In all but the second curve there is a visual symmetry—the top right and bottom left corners of the curve are (ogival) reflections of one another (i.e., doubly reflecting the curve about a central horizontal axis and about a central vertical axis reproduces the curve). This visual symmetry marks symmetric distributional shape. The χ^2_3 quantile function of Figure 3(ii) lacks such symmetry and is positively skewed—as steeper slopes indicate greater spread of density, the shape shows the upper tail to be spread out and the lower to be relatively compact. In general, a positively skewed density has a smile-shaped curve, a negatively skewed density a frown. Tail weights (i.e., probability mass) can also be compared easily—the quantile functions in (i) and (iii) are both symmetric and unimodal but the tails in (iii) are longer than are those in (i). Finally, the quantile function of (v) is without curvature and hence without variation in spread (slope), the mark of a uniform distribution.

Similar examination of the estimated quantile functions, $\widehat{Q}_Y(p)$, of Figure 3's second row tells much the same story about the observed distribution of the data. The first three datasets are unimodal, the first and third are also symmetric with the third having longer tails than the first; the second is positively skewed; the fourth is bimodal and moderately symmetric; the fifth seems to be symmetric and spread fairly evenly across the range of the data. While these descriptions apply to the observed data through $\widehat{Q}_Y(p)$, as descriptions of $Q_Y(p)$ they remain inferences subject to the usual vagaries of sampling variability.

As Figure 3 demonstrates, many points of agreement or disagreement between the shapes of distributions can be revealed by simple visual comparison of their quantile functions.

2.2 Quantile–Quantile Plots

Rather than rely on comparisons across plots, two different quantile functions, say $Q_X(p)$ and $Q_Y(p)$, are more effectively compared by examining the parametric curve $(Q_X(p), Q_Y(p))$ for $p \in (0, 1)$. The plot of these coordinates is called a *quantile–quantile plot* or *qqplot* for short.

Just as curvature in the quantile function means a local change in the spread and hence density, curvature in a qqplot indicates a local change in the relative density of one distribution to the other. In fact, quantile plots as in Figure 3 are special cases of a qqplot. When $X \sim U(0, 1)$, $Q_X(p) = p$, and the corresponding qqplot of $(Q_X(p), Q_Y(p))$ is the same as the quantile plot $(p, Q_Y(p))$.

As with plots of quantile functions, the curve shapes of quantile–quantile plots can be interpreted when the horizontal coordinate is determined by any other quantile function $Q_X(p)$, not just $Q_X(p) = p$.

2.2.1 No Curvature

If the coordinates lie on a straight line, then

$$Q_Y(p) = a + b \times Q_X(p) \qquad (1)$$

for some a and some $b > 0$ and it follows that the distributions are identical up to a location and scale transformation. That is, $\frac{Y-a}{b}$ and X have identical distributions—the distributions of X and of Y have the same shape. Were one to know a, b, and $Q_X(p)$, then the value of $Q_Y(p)$ could be easily found for any value of p.

Based on experience with data such as those appearing in Figure 1(a) and 1(b), Galton (1899) proposed to take advantage of this relationship by plotting the ordered observations $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$ versus the corresponding theoretical quantiles from a standard normal distribution—that is, plotting $(\Phi^{-1}(p), \widehat{Q}_Y(p))$ when $\Phi(x)$ is the standard normal or Gaussian distribution function (see Figure 1(c) for $\Phi^{-1}(p)$). Drawing a straight line on this plot not only provides estimates of location and scale (viz., a and b, respectively) but also provides a more accurate means of estimating percentiles (particularly in the tails) than would direct use of Galton's sample ogive.

Galton's normal qqplot was put to great effect for this purpose years later by civil engineers. Hazen (1914) showed how it could be used to better understand the storage requirements for water reservoirs and the natural variation in the flow of water streams that fed them. Plots such as that of Figure 4(a) were easy to construct and clearly showed the variation in water volumes over many years. Calculations were simplified, various important percentiles easily estimated (e.g., to determine what was meant by a dry or a wet year), and the pattern of different reservoirs or streams could be compared. Qqplots would be published to present the entire dataset and all the essential information that it contained (e.g., Hazen 1914; Whipple 1916a, 1916b).

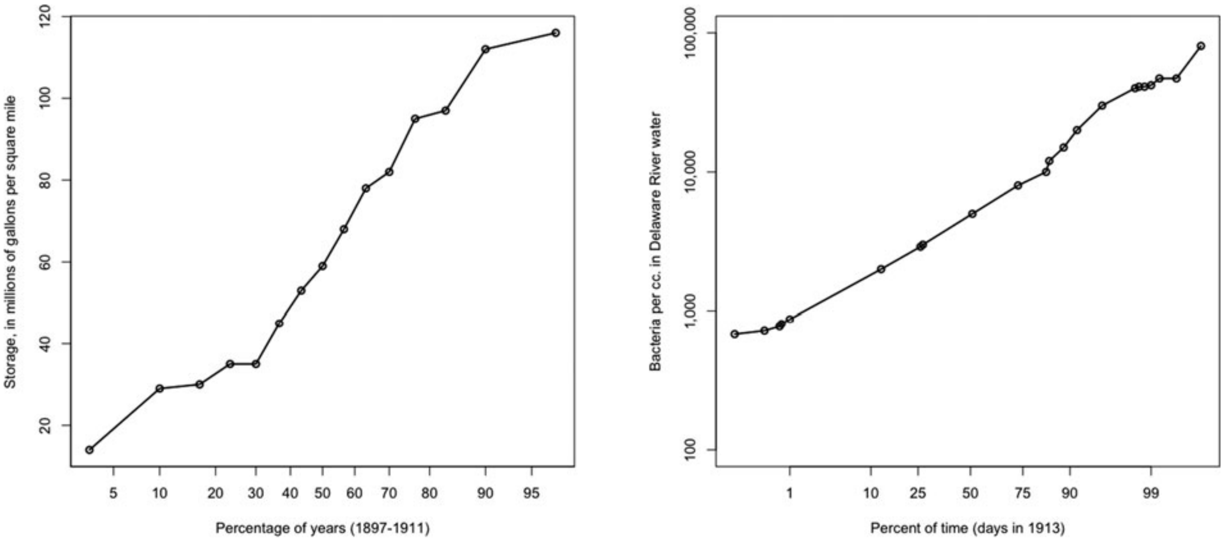
Hazen also noted that oftentimes the data did not follow a straight line as might be expected but rather a "skewed curve" (pp. 626–632 of Fuller 1914). His solution was to plot the logarithms of the data against the normal quantiles as in Figure 4(b)—this would straighten the plot and allow the calculations to be made using the line on the log scale.

Hazen's great innovation over Galton was the construction and printing of "probability paper," paper that had vertical grid lines placed at horizontal locations according to selected normal quantiles. As usual, horizontal grid lines were placed equispaced so that the data values could be plotted naturally. This allowed anyone to easily plot their data simply by knowing their order without any resort to probability tables—*the chance model was built into the technology*. Hazen also had a second grid arrangement, "log probability paper," which additionally had the horizontal grid lines placed according to a logarithmic scale. Using such paper, Whipple (1916b) was able to easily construct the plot of Figure 4(b) without resort to any tables. Both arrangements were quickly taken up by engineers and were the standard means of constructing normal and log-normal qqplots throughout much of the twentieth century, fading in use as software to construct them became more widely available in the 1970s, 1980s, and beyond.

As Galton (1885a) first observed, detecting departure from a straight line is much easier than detecting departure from an ogival curve. With probability paper, any data analyst could quickly plot their data and assess its normality.

2.2.2 Curvature Present

A quantile–quantile plot can reveal much about how one distribution differs from another. A straight line indicates that



(a) Normal qqplot

(b) Log-scaled normal qqplot

Figure 4. Early quantile–quantile plots: (a) Water storage at the Wachusett Reservoir in Massachusetts 1897–1911, Hazen (1914). (b) Bacteria from Delaware River water entering the Torresdale Filter of the Philadelphia water supply 1913, Whipple (1916a, b).

the distributions have identical shape, though they may differ in location and scale. If not a straight line, then as with a quantile plot, the shape of the curve indicates how the distributional shapes differ.

For example, the quantile plots of Figure 3 may be reinterpreted as qqplots where the horizontal quantiles are simply those of a uniform distribution (since for $X \sim U(0, 1)$, $Q_X(p) = p$). The earlier descriptions of the shape of $Q_Y(p)$ now describe how the distribution of Y compares in shape to that of a uniform (or rectangular) distribution. Only the two plots of Figure 3(v) show a straight line (a uniform looks uniform in shape); all others show curvature and hence nonuniform distributional shape.

Figure 5 repeats the figures of Figure 3 but now with horizontal quantiles from the standard normal, $Q_X(p) = \Phi^{-1}(p)$. In the first row, a straight line appears in (i) but not in (v), demonstrating that the distribution of Y in (i) has shape identical to that of an $N(0, 1)$ but that of Y in (v) does not. (The latter, having $U(0, 1)$ quantiles on the vertical axis, is easily recognized as the $N(0, 1)$ cumulative distribution function.) Comments with respect to symmetry, number of modes, and tail weights that earlier applied to Figure 3 also apply here except that the comparisons are now being made between the distribution of Y and that of the standard normal. This change can be seen by the relative straightening of plots (ii), (iii), and (iv) in the first row of Figure 5 compared to the same of Figure 3—in these

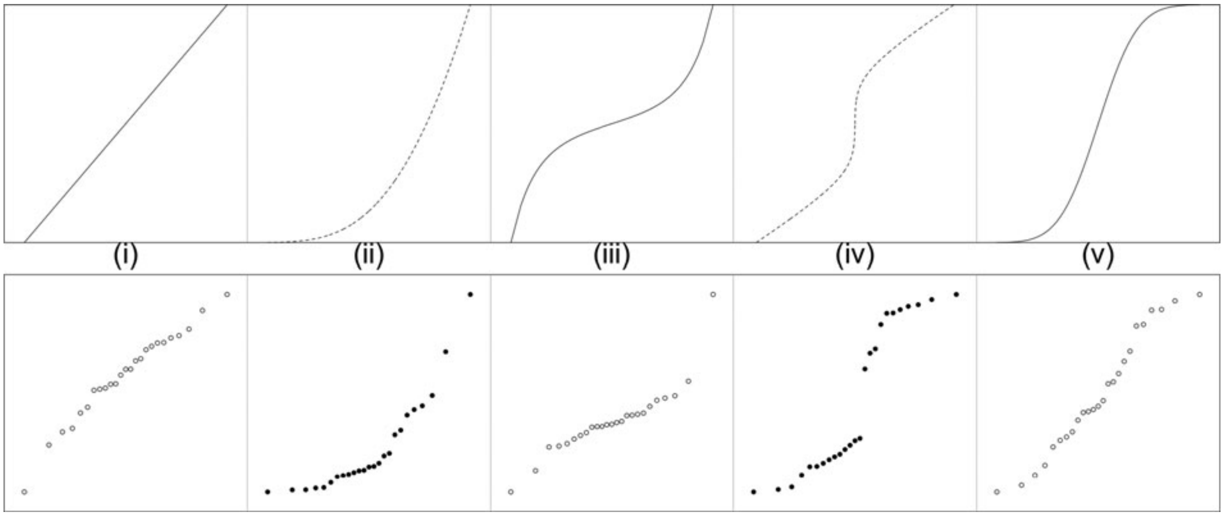


Figure 5. Normal quantile–quantile plots: $(Q_X(p), Q_Y(p))$ with $Q_X(p) = \Phi^{-1}(p)$. The first row has $Q_Y(p)$, the second row has $\widehat{Q}_Y(p)$ for samples of 25 from the same distributions when: (i) $Y \sim N(0, 1)$; (ii) $Y \sim \chi^2_3$; (iii) $Y \sim t_3$; (iv) $Y \sim \frac{1}{2}N(0, 1) + \frac{1}{2}N(6, 1)$; and (v) $Y \sim U(0, 1)$.

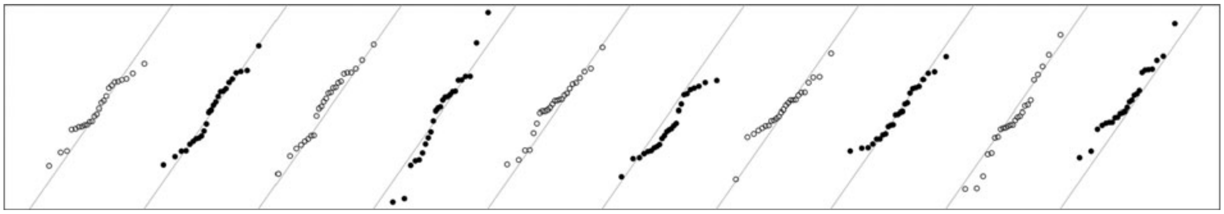


Figure 6. Following Daniel and Wood (1980, pp. 33–43), quantile–quantile plots for 10 independent samples of size 25 randomly generated from $N(0, 1)$. Test distribution is standard Gaussian as well. Gray diagonal lines are the $y = x$ line in each quantile–quantile plot.

cases there is greater agreement between the distribution of Y and the standard normal than there is between Y and a uniform distribution.

2.2.3 *Ambiguity of Sample Qqplots*

The qqplots of the second row of Figure 5 tell essentially the same story as those of the first row, having very nearly the same shapes as the theoretical curves above them. However, where the stories of the first row were definitive, the same stories associated with second row are somewhat more ambiguous because they are based on sample, or estimated, quantiles $\widehat{Q}_Y(p)$ and not on the exact quantiles $Q_Y(p)$. Each new sample will produce a new $\widehat{Q}_Y(p)$ and these will differ somewhat from each other, each suggesting a possibly different story about $Q_Y(p)$.

For example, Figure 5(i) is very nearly a straight line, but not quite. These data are known to be generated from a normal distribution so perhaps such slight departure is to be expected from a sample. Conversely, those of Figure 5(ii–iv) are not from normally generated data and show more obvious departures from a straight line. But are these departures large enough to make the case against normality? The departure from a straight line seen in Figure 5(v) certainly is not as large as in Figure 5(ii–iv). If we did not know how the data were actually generated, how confident would we be about the relative significance of the departures we observe? In real situations such as that of Hazen (1914) and Whipple (1916b), a judgment must be made on whether the plots of Figure 4(a) and (b) support (or refute) the normal and log-normal distributions respectively.

Unfortunately, as Figure 6 illustrates, even when the observed values from Y are known to have been generated from a standard distribution there can be considerable variability in the display from one generated sample to another. Perceived departures from the straight line can be over-interpreted unless the viewer is somewhat experienced in viewing plots as they occur when the null hypothesis holds. Ever since at least Daniel and Wood (1980) students of statistics have been recommended to gain that experience by calibrating themselves against such simulated data.

There are some substantive difficulties with this recommendation. First, it requires some self-discipline to train oneself on many artificial and hence uninteresting plots. Second, such an investment would only pay those who were regularly viewing quantile–quantile plots in application, for others it would be easily lost from long-term memory. Third, the set of such "null" quantile–quantile configurations can be immense and changes with sample size. Finally, the variety of null configurations can change dramatically with the test distribution—for example, training only on normal data may not transfer to other test distributions. A better solution would be to incorporate the training information in the plot itself.

3. SELF-CALIBRATION

To keep things simple, suppose the qqplot has been constructed from n points $(q_i, y_{(i)})$, where $y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)}$ are the ordered observed values of some sample and $q_1 \le q_2 \le \cdots \le q_n$ are the corresponding quantiles $q_i = Q_X(p_i)$ from a specified test distribution $F_X(x)$. To be concrete, for the moment take $p_i = (i - \frac{1}{2})/n$ as in Hazen (1914) (in `qqtest` other default values are taken for p_i depending on the test distribution following the recommendations of Cunnane (1978); the user may also supply their own).

Because $Q_X(p)$ is given, an ordered sample can always be generated from the test distribution by first randomly generating n values $p_i \sim U(0, 1)$ independently for $i = 1, \ldots, n$, determining $x_i = Q_X(p_i)$ for each i and then sorting the values $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. A plot of the pairs $(q_i, x_{(i)})$ provides an *exemplar* or *null configuration* against which the qqplot of $(q_i, y_{(i)})$ may be compared.

Looking at many such exemplars as in Figure 6 (where $F_X(x) = \Phi(x)$) provides a sense of their variability and also allows some visual assessment as to how much the observed qqplot of points $(q_i, y_{(i)})$ resembles that of a typical qqplot for samples from the test distribution. However, this comparison could be made much more directly if the plot of $(q_i, y_{(i)})$ could simply be overlaid with that of $(q_i, x_{(i)})$.

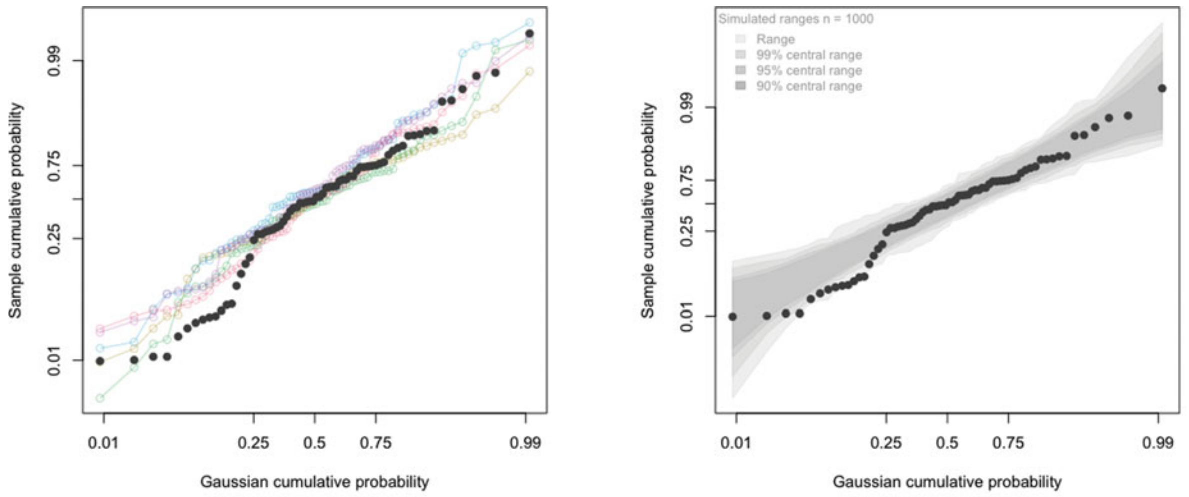
Before such overlaying is possible however, the location and scale of $(q_i, x_{(i)})$ must be matched to that of the observed data $(q_i, y_{(i)})$. This is easily done by fitting a line

$$y = a + b \times q$$

to the values of $(q_i, y_{(i)})$. Since we expect such a plot to often have outliers, a robust regression is used to get suitable values for a and b (`qqtest` employs a high breakdown point robust regression via `lmRob` of the R package `robust`, Wang et al. (2014)—high breakdown is especially important for small samples). With a and b in hand, the x values can then be relocated and scaled to match that of the y's thus allowing the essential shapes of the two curves to be compared. That is, the x values are generated as

$$x_i = a + b \times Q_X(p_i)$$

for $p_i \sim U(0, 1)$ for $i = 1, \ldots, n$.



(a) Exemplars: 5 overlaid from $\Phi(x)$

`qqtest(precip, nexemplars=5, envelope=FALSE)`

(b) Percentile envelopes: 1,000 samples from $\Phi(x)$

`qqtest(precip)`

Figure 7. Quantile–quantile test plots for the average amount of precipitation (rainfall) in inches for each of 70 United States (and Puerto Rico) cities in 1975. Test distribution is standard Gaussian.

Several such exemplars could easily be overlaid on the same plot as the observed data as shown in Figure 7(a) where five exemplars are laid over the normal qqplot of the `precip` data from R. Each exemplar is an independently generated sample from the test distribution, rescaled, and relocated as described above. Both the individual points $(q_i, x_{(i)})$ and the line segments joining them are plotted for each exemplar—these produce the colored trails shown. Each trail is plotted using a transparent color so that colors combine where overlaid (i.e., via alpha-blending).

Alternatively, a great many exemplars could be generated and only a summary of their information content presented. Each exemplar provides a randomly generated ordered value $x_{(i)}$ so generating a great many such values will give a reasonable approximation to the distribution of $x_{(i)}$ for any i.

Figure 7(b) shows the summary results for 1000 generated samples from the normal distribution. Central 90%, 95%, and 99% bands are formed at each q_i from the corresponding sample of 1000 values of $x_{(i)}$. These values, as well as the minimum and maximum of each set of 1000, are joined together to form envelopes, each providing an empirical point-wise confidence region for exemplar quantile–quantile curves. The envelopes are each drawn with the same transparent gray so that alpha blending ensures that the envelopes become darker as the confidence levels increase (since the envelopes are nested). The crisp edges seen are a natural artifact of the well-known "Mach effect" optical illusion. As the code below Figure 7(b) shows, this plot is the default plot produced by the `qqtest(·)` function. That below Figure 7(a) shows how to produce one containing only exemplars.

As can be seen from either plot of Figure 7, the `precip` data points (in black) indicate a bimodal shape with modes separating just before the first quartile. In Figure 7(b), we see that there is strong (point-wise) evidence against the hypothesis that the `precip` data come from a normal distribution—just before the first quartile the `precip` points appear at, or just outside, the range of 1000 samples generated from a normal distribution. Figure 7(a) tells the same story but is based on only comparison with five exemplars so the evidence is not as strong. The advantage that Figure 7(a) has over that of (b) is that the whole trail of each exemplar can be seen. This allows one to make an assessment based on the entire shape of the trails rather than simply on the point-wise limits—the shape of the `precip` quantile–quantile curve is very different from any of the five exemplars. The bimodality seen in the `precip` data appears to be real rather than being a spurious artifact of sampling.

3.1 Null Configurations

Applying `qqtest` to each of the normally generated samples of Figure 6 gives some sense of the default behavior of the function when the null hypothesis holds (in this case that of normality). As Figure 8, shows, in no case is there strong evidence against the hypothesis of normality. The greatest evidence would seem to be the seventh sample plot of Figure 8 where the lowest point appears inside the 99% envelope and outside the 95%. This corresponds to a *point-wise* significance level between 1% and 5%. As there are 25 points in each plot, there are really 25 point-wise tests being applied. Taking into account the problem of multiple tests, the significance level is actually much larger than 5% and the plot provides little to no evidence against normality.

3.2 Nonnull Configurations

Applying `qqtest` to each of the generated samples of Figure 5 will similarly give some sense of the default behavior of the function under various alternatives to the null hypothesis, namely, that the test distribution (in this case the normal) generated the data. The results appear in Figure 9.

At leftmost is a null configuration where the data were generated according to the test distribution and the plot gives no evidence against the hypothesis of normality. In contrast,

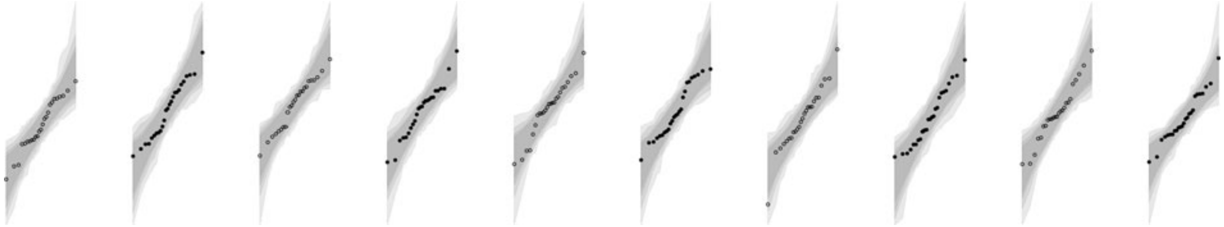


Figure 8. `qqtest(·)` applied to each of the samples from Figure 6.

the next three plots (ii)–(iv) all show strong evidence against the hypothesis of normality—(ii) has many points far outside the range of 1000 samples from the test distribution, (iii) has one such point, and (iv) has several points at the edge of the outermost envelopes and its bi-modal curve is easily distinguished in a plot overlaid with exemplars, as was done with the `precip` data of Figure 7(a). From left to right, the strength of evidence against the hypothesis declines until in (v) there appears to be no evidence against the hypothesis of normality even though the data were indeed generated from a $U(0, 1)$ distribution. The data of (v) do show slightly shorter tails on both sides, which might be further investigated via exemplars to compare curve shapes (e.g., as in Figure 7(a)).

4. EXAMPLES

In the more than 100 years since its introduction, the quantile–quantile plot has been recommended for visual assessment of a wide variety of statistical data. This includes, for example, experimental effects, residual analysis and outlier detection, and the direct comparison of any two sets of univariate data (e.g., Daniel 1959; Wilk and Gnanadesikan 1968; Gnanadesikan 1977; Daniel and Wood 1980; Chambers et al. 1983). The quantile–quantile plot can be used on any sample, or pair of samples, of univariate data—however large any sample might be. In this section, the self-calibrating qqplot is illustrated on a number of such examples.

4.1 Revisiting the Earliest Qqplots

Figure 10 shows the result of applying `qqtest` to Galton's data of Figure 1(a) and 1(b) and to Hazen's of Figure 4(a). With the envelopes the plots of Figure 10 have much more to say about the various datasets than did the original qqplots.

While there are only a few points in each of Figure 10(a) and 10(b), each point is a sample percentile determined from a sample of a great many more points ($n = 519$ and $n = 775$, respectively). It is these very large original sample sizes that bring about the relatively narrow envelopes seen here. Because of these very large sample sizes, the single point appearing at the edge, or just outside, of the range envelope in Figure 10(a) indicates strong evidence against the hypothesis of normality for Galton's pull strength data. The pull strength data seem to have a much longer right tail than would be produced by the vast majority of samples from a normal distribution. Conversely, Galton's sitting height data of Figure 10 show no evidence against the hypothesis of normality. More worrying perhaps is that its sample quantile–quantile curve appears to be suspiciously close to perfect.

The envelopes on Hazen's water reservoir storage data in Figure 10(c) are much wider because they are based only on a sample of size $n = 15$. There would seem to be no evidence against the hypothesis of normality here either. Note however that at the left most side, a large fraction of even the 90% central envelope is below zero. This in itself provides considerable evidence against normality for the data—it is not possible to have less than zero millions of gallons of water. In this case, the hard boundary is evident and indicates that the left tail of the storage distribution is necessarily much shorter than a normal. Fortunately, Hazen (1914) was more interested in using the quantile–quantile plot to estimate various percentiles for water storage than its fidelity to the normal distribution.

It is worth noting that the construction of these plots differs slightly from one another. For Hazen's data shown in Figure 10(c), the entire sample is available and the plot is constructed as described previously. For Galton's data, a slightly different though statistically equivalent construction is used by `qqtest`.

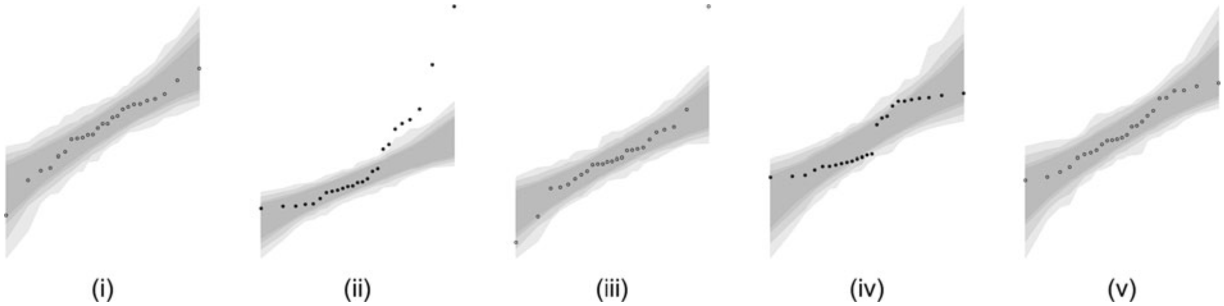
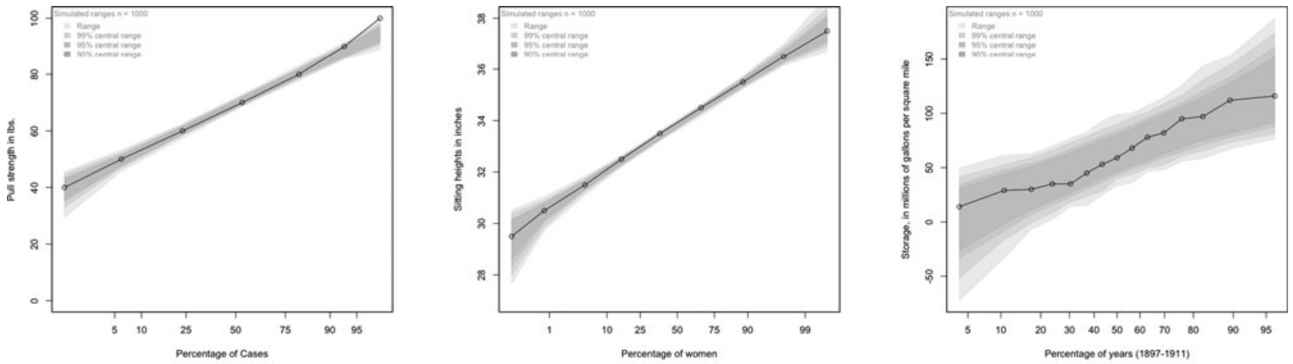


Figure 9. `qqtest(y)` applied to each of the samples from Figure 5.



(a) Pull strength `qqtest(y, np=519, p=pVals,...)`

(b) Sitting heights `qqtest(y, np=775, p=pVals,...)`

(c) Reservoir storage `qqtest(y, ...)`

Figure 10. Normal (default) `qqtest` applied to the datasets of Galton (1889), Galton (1885a), and Hazen (1914), respectively.

Neither dataset of Galton includes the original data points y_i but only the sample values at some selected percentiles $p_1, \ldots, p_k$, say, and the total sample size, n ($n = 519$ and $n = 775$, respectively, in Galton's two cases). In this case, `qqtest` generates an individual sample for the order statistic $x_{(i)}$ corresponding to each percentile p. This is done by exploiting the fact that ith-ordered value of a $U(0, 1)$ random variable is distributed as a beta random variable. That is for a sample of size n from a $U(0, 1)$, the ith largest value has distribution $U_{(i)} \sim \text{Beta}(i, n + 1 - i)$. An observed value $x_{(i)}$ is simply generated using the relationship $X_{(i)} = Q_X(U_{(i)})$. Taking $i = n \times p$ for each selected percentile p, `qqtest` need only have the values of n (as `np`) and p (as the vector `p`) to produce the envelopes.

4.2 Choosing a Scale—The Log-Normal

Figure 11 shows the result of applying `qqtest` to Whipple's data of Figure 4(b). Figure 11(a) shows the raw data plotted against the quantiles from a log-normal distribution. The points at the extremes of the envelope (near the 95th percentile) are evidence against the hypothesis that the bacterial count is log-normal. This plot has the advantage of preserving the original linear scale on the vertical axis but, as the envelopes show, there is considerable variation in the sample quantiles, particularly at the high end and the data are severely compressed for all percentiles below 80.

A statistically equivalent approach is that of Hazen (1914), namely, to take the logarithm of the data and compare that to normal quantiles. Figure 11(b) shows the result of applying `qqtest` in this case. The vertical axis is now a log-scale, the horizontal quantiles are more spread out, and the vertical spread of the envelopes is less variable. Several points are now seen to be at the edge of the envelopes providing very strong evidence against the hypothesis of a log-normal distribution.

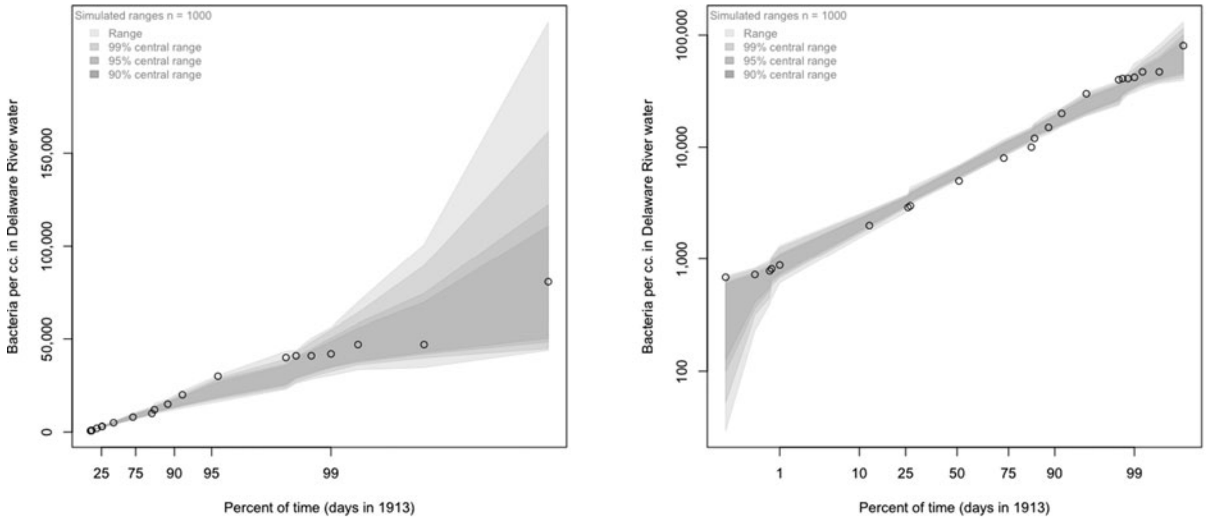
4.3 Factorial Effects—The Half-Normal

In the early days of its development, penicillin was produced by growing the fungus *penicillium chrysogenum* in a nutrient medium including a mix of ingredients in various concentrations. Experiments were conducted to try to determine which ingredients, and in what concentrations, would optimize production. Because of the possible interdependence between the various ingredients and the consequent effect on production, factorial designs were used extensively to tease out important main effects and significant interactions.

Davies (1956) presented the design, data, and analysis for one such experiment (viz., Example 9.2, p. 383ff, p. 416ff). The design was a 2^{5-0} factorial, with the five two-level factors under investigation being [A] corn steep liquor (strength 2% and 3%), [B] lactose (2% and 3%), [C] precursor (0% and 0.05%), [D] sodium nitrate (0% and 0.3%), and [E] glucose (0% and 0.5%). The response was the logarithm of the measured yield of penicillin. All 32 combinations of factors were used with no replication and 16 runs executed in 1 week of production and 16 in the next—the design confounded the two blocks with the ABCDE interaction term. Three- and four-way interaction terms were used to estimate the error variance for standard testing of effects. A standard analysis shows three main effects, A, C, and E, and one two-factor interaction, CE, to be statistically significant, the last of these only at a 5% level.

There is a problem with this analysis though. Standard significance testing presumes that the hypothesis to be tested has been determined before examining the data, whereas factorial designs are more typically used to test hypotheses that are determined after seeing the data, namely, those hypotheses that match the larger observed effects. Unfortunately, as Daniel (1959) pointed out, for a 32 run 2^{5-0} experiment *having no real effects*, a standard test will find the largest effect to be significant at the 5% level about *half the time!* Any reliable assessment must take some account of this post hoc hypothesis selection or there will always be numerous false positives.

This assessment is considerably improved by use of the half-normal or Daniel plot introduced by Daniel (1959). Here, the unsigned and ordered estimated effects are plotted against the corresponding quantiles of a half-normal distribution (i.e., $X = |Z|$ where $Z \sim N(0, 1)$)—a qqplot with the half-normal as test distribution. As such, departures from a straight line indicate departure from the hypothesis of null effects. Now the importance of an effect is associated more with the size of its



(a) Log-normal qqtest

`qqtest(count, dist="log-normal", np=365, ...)`

(b) Log-scaled normal qqtest

`qqtest(log(count,10), dist="normal", np=365, ...)`

Figure 11. `qqtest` applied to bacteria count data of Whipple (1916b).

departure from the line of data points than with its absolute magnitude. How large a departure from the straight line should be considered significant still depends on the order position of the effect in the plot.

Figure 12(a) shows the half-normal qqplot produced by `qqtest` for these data. As can be easily seen, three main effects (C, A, and E) are clearly significant, each one being outside the range of 1000 replicates. The interaction term CE however is judged not to be statistically significant—Davies (1956, pp. 386–387, 416–417) cautiously concluded that it was, though it would seem largely on the strength of a priori information about the relationship.

It is also clear from Figure 12 that the simulated bands increase in width the larger is the observed effect being considered. This is easily seen mathematically as well. The random variable Y is related to $U \sim U(0, 1)$ through the probability integral transform $Y = Q_Y(U)$. Taking a first-order Taylor approximation $Y \approx Q_Y(p) + (U - p)Q_Y{}'(p)$, the random error for the ith largest value corresponding to p_i (e.g., $p_i = (i - 0.5)/2$) can be written as

$$Y_{(i)} - Q_Y(p_i) \approx \left[Q_Y{}'(p_i)\right] \times (U_{(i)} - p_i). \qquad (2)$$

Typically, the dominant term here will be the derivative $Q_Y{}'(p_i)$, the slope of the quantile function at p_i. One need only examine the right half of a normal ogive (Figure 3(i), first row) to appreciate how large the magnitude of this slope can be for a half-normal distribution. Clearly, those effects having p_i nearer 1 on a half-normal plot must demonstrate a greater visual departure (proportional to $Q_Y{}'(p_i)$) before being marked as statistically significant. For other qqplots, the regions of large squared error can also be read from the squared slopes of the quantile function (e.g., Figure 3(ii–v), first row).

This was not lost on Daniel (1959) who recommended adding "guardrails" to the plot corresponding to various one-sided critical values (based on the simulated and mathematical results of Birnbaum (1959) as well as his own simulations). Figure 12(b) shows the simulated central percentiles as guardrails for some values used by Daniel (1959) (his one-sided $\alpha = 0.4, 0.2,$ and 0.05 have central percentiles of $0.2, 0.8,$ and 0.9). For some experimental purposes, Daniel (1959) recommended using $\alpha = 0.4$ (central percentile 0.2) and even larger. If one is trying to find all real effects then keeping more false positives would ensure that fewer false negatives would be lost—the hope would be that further experimentation would tease out the real from false effects. For example, in Figure 12(b) $\alpha = 0.2$ would have marked CE as significant and Daniel's choice of $\alpha = 0.4$ would have marked many more effects smaller than CE as possibly worth retaining for further experimentation. Practitioners have since suggested many alternative ways to assess the significance of outlying points on a half-normal plot used for unreplicated experiments (e.g., see Hamada and Balakrishnan 1998 for a review).

As with every qqplot there is considerable information that can be seen from the configuration of points beyond which of them are far from a straight line. This is particularly true in the case of half-normal plots for the contrasts from a factorial design. Daniel (1959) discussed these at some length for this context.

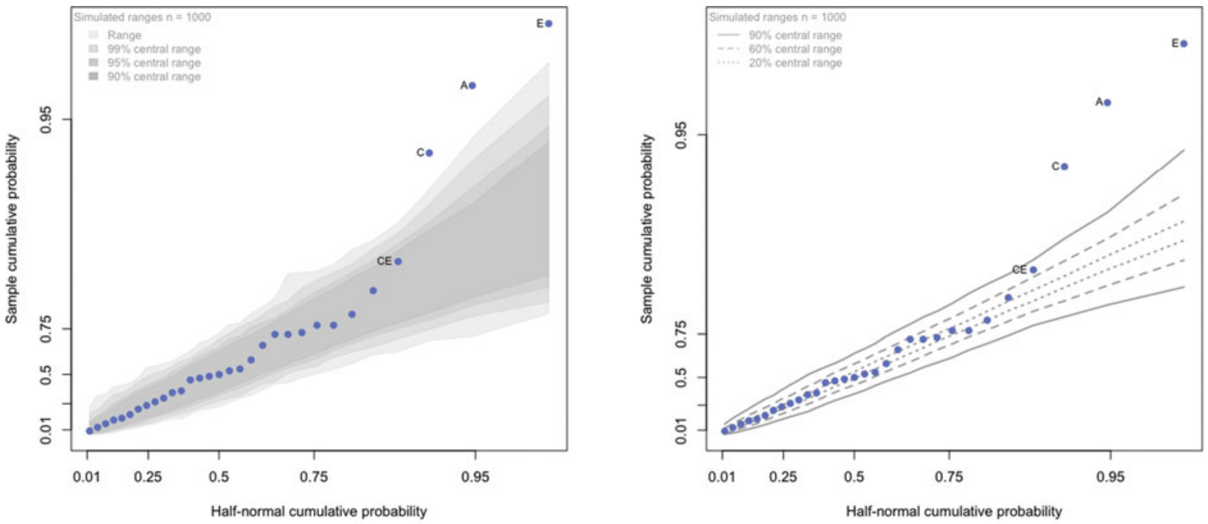
In industrial experimentation, half-normal or Daniel plots are standard practice but have recently been criticized in Lenth (2015) for being too difficult for routine interpretation. The routine interpretation being criticized (departure from a straight line) is somewhat distant from the careful interpretation and caveats described in Daniel (1959) or made obvious by Equation (2).

In contrast and in keeping with the original philosophy of Daniel (1959), the addition of the envelopes or selected guardrails as in Figure 12(a) or 12(b) allows a different and much more reliable routine interpretation, one which clearly and correctly identifies important effects (see applying `qqtest` to the examples of Lenth 2015). Moreover, the output of `qqtest` still allows the deeper interpretations available of the configuration as given in Daniel (1959).

4.4 Statistical Process Improvement—Monitoring $\bar{x}$ and s

It is common practice in process improvement to measure important characteristics of individual units as they are produced and follow how these measurements change over time—process monitoring. To this end, many charts have been developed and put to use with good effect to improve product quality.

Two such charts are the $\bar{x}$ chart and the s chart that follow estimates of the process mean μ and standard deviation σ over time. First a group of m items $i = 1, \ldots, m$ are selected at a time from the process and a measurement x_i taken on each item i. For each group, $\bar{x} = \sum_{i=1}^{m} x_i/m$ and $s = \sqrt{\sum(x_i - \bar{x})^2/(m-1)}$ are cal-



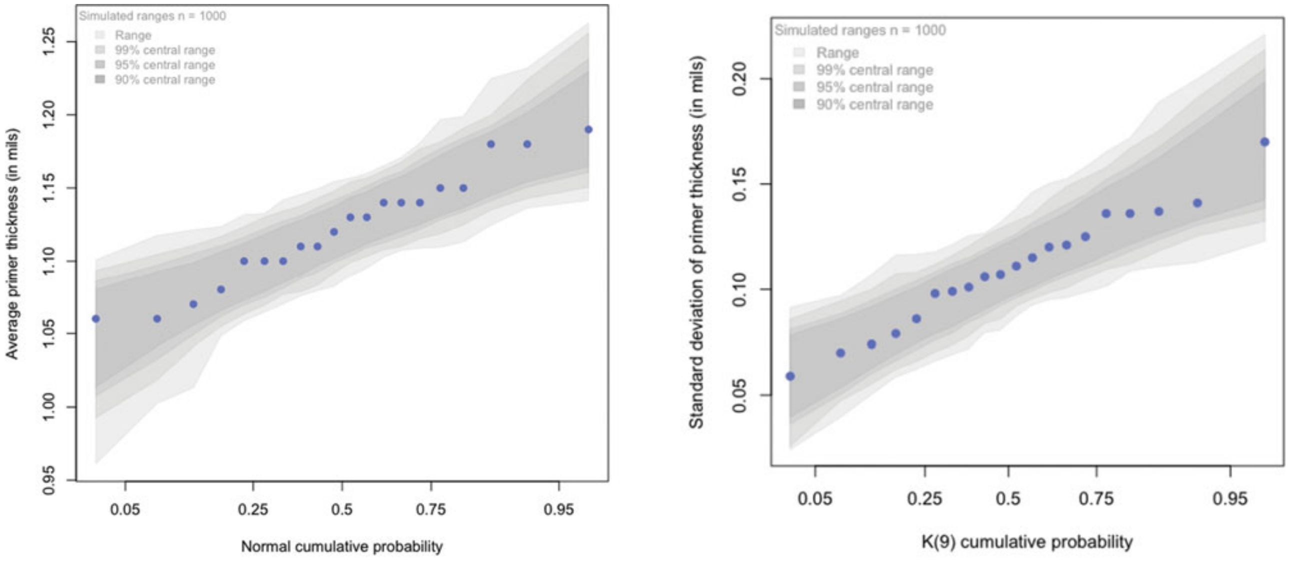
(a) Half-normal qqtest

```
qqtest(y, dist="half-normal")
```

(b) Guardrails instead of envelope

```
qqtest(y, dist="half-normal",
          envelope=FALSE, drawPercentiles=TRUE,
            centralPercents=c(0.2, 0.6, 0.9))
```

Figure 12. The half-normal plot. Thirty-one unsigned contrasts from a 2^{5-0} factorial experiment on penicillin production used by Daniel (1959) from Davies (1956).



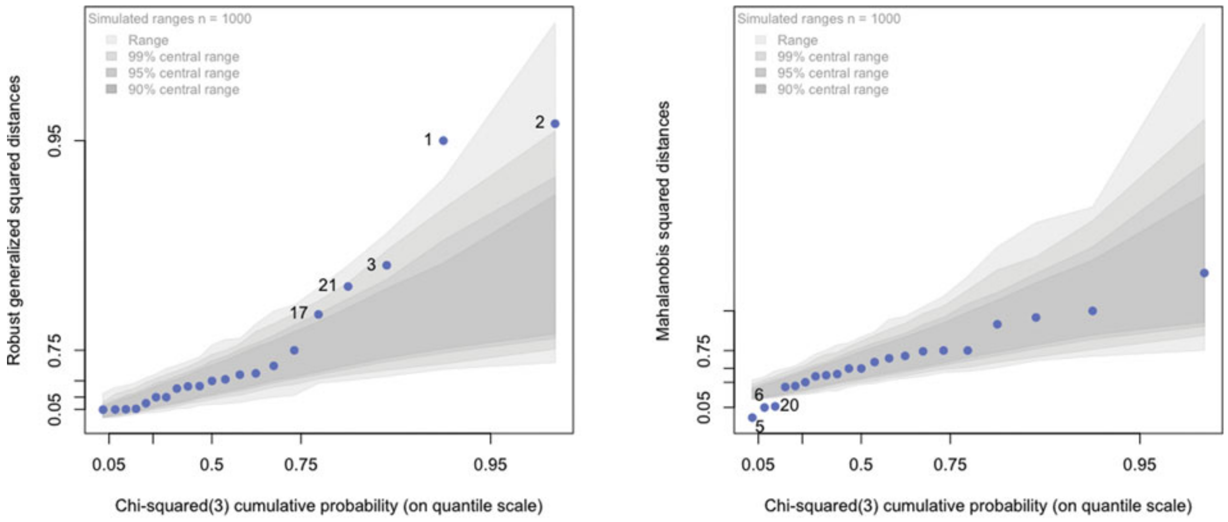
(a) $\bar{x} \sim N(\mu, \sigma^2/10)$

```
qqtest(xbar, ...)
```

(b) $s/\sigma \sim K_9$

```
qqtest(s, dist="kay", df=9, ...)
```

Figure 13. `qqtest` applied to $\bar{x}$ and s from page 64 of AIAG (1992).



(a) Robust estimates (b) Standard estimates

```
qqtest(distances, dist="chi-squared", df=3, ...)
qqtest(distances, qfunction=function(x){qgamma(x, shape=3/2)}, ...)
```

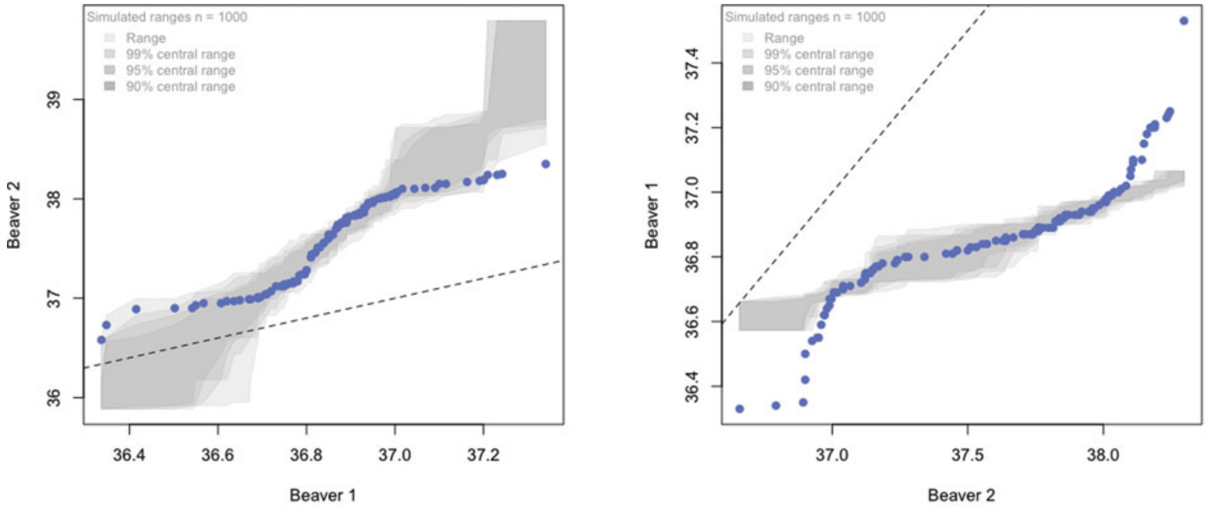
Figure 14. `qqtest` applied to the three independent regressor variables of the stackloss data from Brownlee (1960, p. 491).

culated. If we let $\bar{x}_t$ and s_t be the values at time t then the $\bar{x}$-chart plots the pairs $(t, \bar{x}_t)$ and the s-chart plots the pairs (t, s_t). Each chart is supplemented by horizontal lines representing "control limits"—points beyond these lines are indicators that some action should be taken to bring the process under control.

The control limits are intended to be about 3 standard deviations beyond the target or central value being measured. The standard deviation in question is that of either $\bar{x}$ or s, depending on the chart, and is determined by assuming that the original x_i's are independently $N(\mu, \sigma^2)$. Unfortunately, any such chart will contain many values of $\bar{x}$ or s, the largest of which will always have higher probability of exceeding a limit than will any one $\bar{x}$ or s value individually. The control charts do not take this into account.

In place of either chart, we might consider an appropriate qqplot as in Figure 13(a) and 13(b).

These values are taken from the training manual AIAG (1992) where the thickness in mils (thousandths of an inch) of primer paint applied to an automotive part is recorded as part of a



```
qqtest(beaver2$temp, dataTest=beaver1$temp, ...)
```
(a) Test = beaver 1 distribution

```
qqtest(beaver1$temp, dataTest=beaver2$temp, ...)
```
(b) Test = beaver 2 distribution

Figure 15. Beaver tales. Comparing distributional shapes. (a) Could the distribution of body temperatures for Beaver 2 have been generated from the same distribution as that of Beaver 1? (b) Or vice versa? Data source: `beaver1` and `beaver2` data from R.

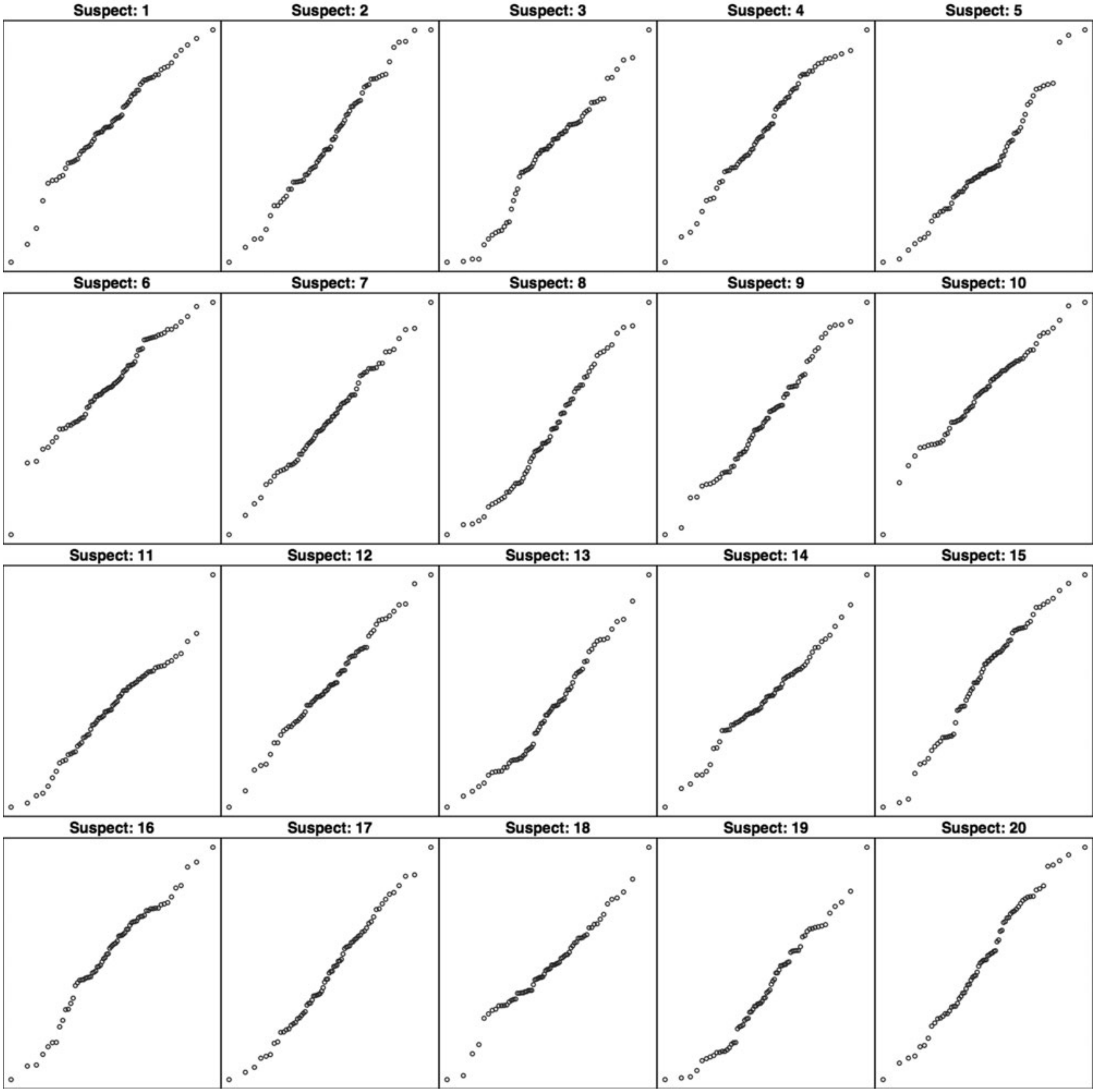


Figure 16. Line-up test for normal qqplots. The observed data are displayed in just one of the 20 plots shown; every other is sample generated from the test distribution. Location of the observed data is suspect number: `log(4.23911582752162e+28, base=27) - 17`.

monitoring study. Ten consecutive pieces were measured twice a day for 10 days. Each value of $\bar{x}_t$ and of s_t is based on 10 measurements modeled as a sample from a $N(\mu, \sigma^2)$ for $t = 1, \ldots, 20$. Under these assumptions $\bar{x}_t \sim N(\mu, \sigma^2/n)$ and $s/\sigma \sim K_9$, the latter being the "kay" distribution on 9 degrees of freedom and which is related to the chi-squared as $K_m = \sqrt{\chi^2_m/m}$. The appropriate arguments to `qqtest` are shown under each plot. As can be seen in either plot, no point is outside the envelope—there is no evidence against the hypothesis that the process is behaving as assumed.

Note that although the time order does not appear in the qqplots of Figure 13, there is no longer any need to calculate control limits. These are automatically calculated by the envelopes—the location and scale calculations implicitly incorporate the estimates of the overall averages and standard deviations, and the envelopes adjust to the relative magnitude of the group being considered. As each new value ($\bar{x}$ or s) is gathered it is added to the qqplot and the envelopes update. The plots adjust automatically for the number of groups and the size of the group. Time order could still be displayed in a visually linked companion plot as needed.

4.5 Multivariate Outlier Detection—Generalized Distances

Probability plots of generalized (elliptically contoured) distances have long been used to help identify outliers in multivariate data (e.g., Wilk and Gnanadesikan 1964; Gnanadesikan and

Kettenring 1972; Gnanadesikan 1977). For multivariate observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, the squared generalized distances are defined to be

$$d_i^2 = (\mathbf{x}_i - \mathbf{c})^T \, \mathbf{D}^{-1} \, (\mathbf{x}_i - \mathbf{c}) \,, \tag{3}$$

where $\mathbf{c} \in \mathbb{R}^p$ is a measure of the center of the data and $\mathbf{D}$ is a symmetric positive definite matrix chosen to capture the dispersion of the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^p$.

For example, if $\mathbf{D} = \mathbf{S}$ and $\mathbf{c} = \bar{\mathbf{x}}$ are the sample covariance matrix and mean vector (i.e., the standard estimates), then d_i^2 is the squared Mahalanobis distance. Large values of d_i^2 should be associated with outlying points. Assuming multivariate normality, the squared distance is approximately χ_p^2 and both Cox (1968) and Healy (1968) suggested identifying outliers by plotting a qqplot of these Mahalanobis squared distances against the quantiles of a χ_p^2 or gamma distribution as in Wilk, Gnanadesikan, and Huyett (1962) (alternatively d_i could be plotted against those of K_p). Unfortunately, if there are outlying points in the sample, these can adversely affect both $\bar{\mathbf{x}}$ and $\mathbf{S}$, pulling the average away from the center and inflating the variability in some directions.

Alternatively, a number of authors (e.g., Gnanadesikan and Kettenring 1972; Gnanadesikan 1977; Campbell 1980; Rousseeuw and Van Zomeren 1990) have suggested using robust estimates of location and covariance matrix. The resulting robust generalized distances should more easily display outlying points when plotted in a χ_p^2 qqplot. Figure 14 shows the χ_3^2 qqplot produced by `qqtest` applied to 21 points in $\mathbb{R}^3$. The data are taken from the `stackloss` dataset first appearing in Brownlee (1960). The data have been much used in robust and exploratory methods with respect to outlier detection (e.g., see Dodge 1996). They consist of measurements from 21 days of operation of an industrial process oxidizing ammonia to nitric acid. The three variables used here correspond to the independent variables "rate of operation of the plant," "the temperature of some cooling water," and a transformed "concentration of acid circulating." The traditional response variable "stack loss" is not used here.

As Figure 14(a) shows, five outlying points can be identified by calling `qqtest` on the robust generalized squared distances (using the robust location and covariance provided by the `covRob` function of R package `robust`—Wang et al. 2014). In order of importance the outliers are seen to be 1, 2, 21, 3, and 17. With the exception of 17, these are the points identified as significant in Rousseeuw and Van Zomeren (1990).

Figure 14(b) shows the same call on the Mahalanobis distances as suggested by Cox (1968) and Healy (1968). The standard estimates $\mathbf{S}$ and $\bar{\mathbf{x}}$ are so poor that the qqplot shows no outliers. Note however that `qqtest` still indicates something is amiss. There now appear to be three *in*liers—5, 6, and 20—whose squared distances by the standard estimates are much smaller than would be expected. This is essentially due to the inflation of the sample variance in some directions due to outliers.

Just above the caption of Figure 14 are two different calls to `qqtest` that would effect the same plot—one that uses the ″`chi-squared`″ value of the `dist` argument with degrees of freedom `df=3`; the other uses the more general `qfunction` argument and passes as its value a function that calculates the quantile of a gamma random variable with appropriate parameter values as in Wilk, Gnanadesikan, and Huyett (1962).

4.6 Comparing Samples

Two samples can be easily compared via qqplots. Empirical distributions are compared by plotting the pairs $(\widehat{Q}_X(p), \widehat{Q}_Y(p))$ for a collection of values of p, or $(x_{(i)}, y_{(i)})$ when sample sizes are equal. If the resulting curve is very nearly a straight line, then the traditional location and scale comparisons can be made via the line of Equation (1)—equal locations or means correspond to a zero intercept, equal scales or standard deviations correspond to a unit slope. Differences in the distributional shape are indicated by the quantile–quantile curve.

Figure 15 illustrates how this would look when comparing two datasets—in this instance the body temperatures of two beavers. In each plot, the $y = x$ line has been added to aid location and scale comparisons. In Figure 15(a), the empirical distribution of the sample temperatures from beaver 1 is used as the test distribution and beaver 2's temperatures are seen to come from a distribution having shorter tails than does beaver 1's. In Figure 15(b), it is beaver 2's sample that gives the test distribution—again beaver 1's temperatures are seen to come from a distribution having longer tails than does beaver 2's.

5. VISUAL SIGNIFICANCE TESTS

While the envelopes of a `qqtest` provide a tool for assessing the evidence against the hypothesis that the data were generated by the test distribution of the horizontal axis, that test is based entirely on many individual point-wise tests. In contrast, overlaying exemplars rather than envelopes (i.e., as in Figure 7(a) vs. (b)) allow the shapes of the individual curves to be compared.

Alternatively, rather than overlay the different exemplars one might instead lay them out spatially in separate but nearby plots. This of course makes comparison more difficult and less reliable as the eye has to traverse much larger distances to evaluate differences. The distinct advantage however is that if carefully and honestly done it can provide some measure of the statistical significance of the observed data's departure from the test distribution.

The protocol is that of a "lineup test" as first formally described by Buja et al. (2009). The metaphor is based on the police lineup of suspects from which a witness must identify the guilty party. In our case, the witness is the data analyst and the suspects an arrangement of quantile–quantile plots of data generated from the test distribution together with the one quantile–quantile plot of the observed data. Unlike the metaphorical witness, however, the analyst must not have seen the quantile–quantile plot of the data. This is essential for any honest determination of statistical significance.

Figure 16 shows the lineup plot for 20 subjects, one of which is the quantile–quantile plot of the real data and the remainder are of datasets of the same size but generated from the test

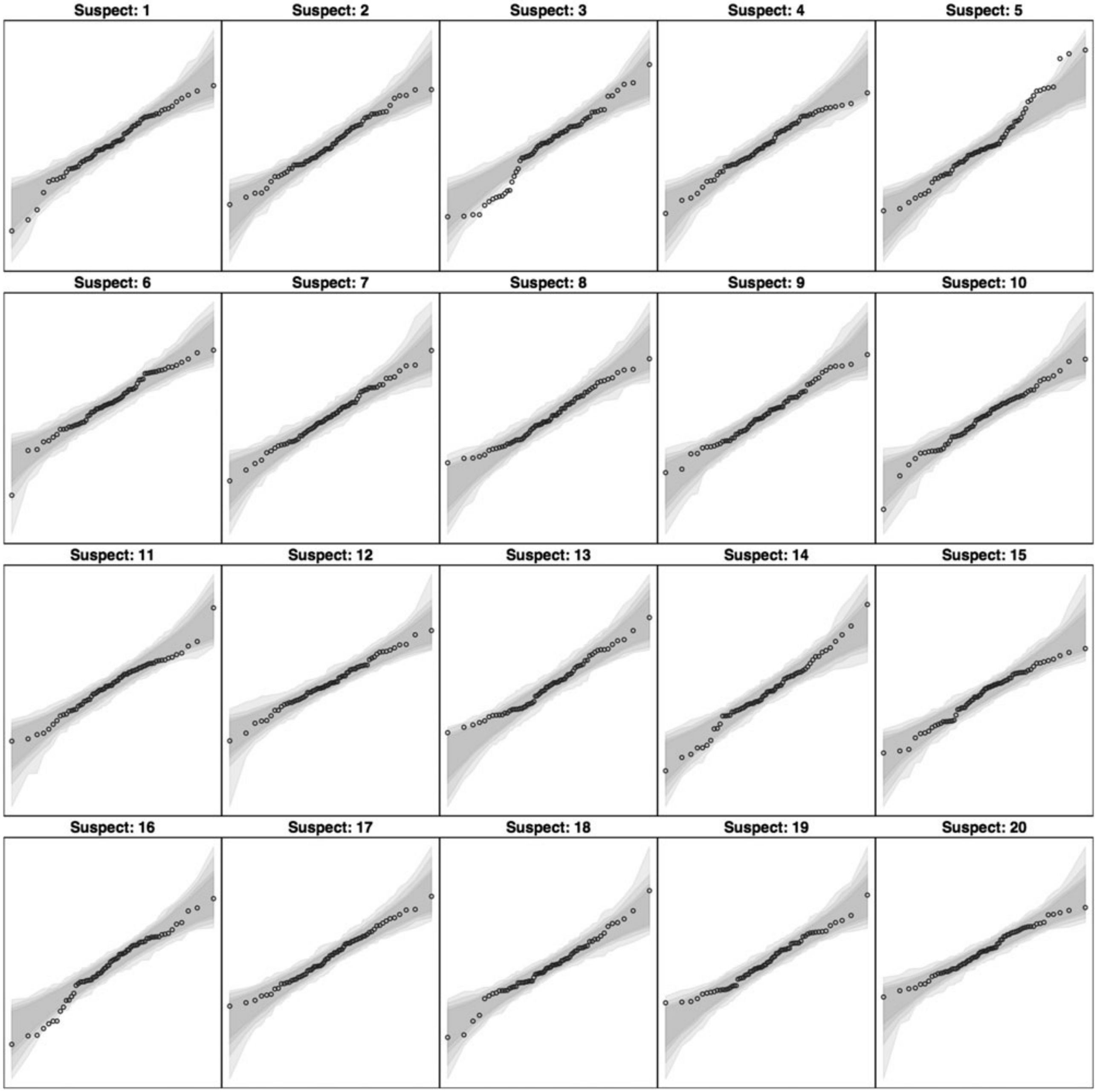


Figure 17. Line-up test for normal qqplots with an envelope. The observed data are displayed in just one of the 20 plots shown; every other is sample generated from the test distribution. Location of the observed data is suspect number: `log(1.75478058540818e+130, base=28) - 87`

distribution. The true location of the real data's quantile–quantile plot is randomly allocated. Out of the 20 plots, the analyst should choose that which gives the greatest indication (in their view) of a dataset that has not come from the test distribution. If the true location/suspect is identified, then the observed significance level is 1/20, otherwise it is greater.

Just as with police lineups, however, this statistical test has the peculiar feature that the test statistic being used is a function both of the particular statistical method (in Figure 16 this is the standard normal quantile–quantile plot) *and* of the particular analyst making the judgment—one hopes that more experienced analysts are statistically more "powerful" than the less experienced.

Similarly, some particular visual presentations might be more powerful than others. Figure 17 is a lineup plot for quantile–quantile plots that include generated envelopes. The observed data are the same as in Figure 16. If the observed data are more often chosen (among different analysts) for quantile–quantile plots with envelopes than it would be without, then this would indicate that adding the envelope increases the discriminatory power of the plot. Whatever the display, however, its power can be substantially diminished when the space available is very small (e.g., one need only compare the visible content of one plot of Figure 17 to that of Figure 7(b) to appreciate the loss).

6. CONCLUDING REMARKS

Quantile–quantile plots have long been valued for their versatility and wide applicability. They are standard practice in applied statistics, useful for model checking, residual analysis, and informal exploratory data analysis. They are however not easily interpreted without practice. One of the great difficulties is that correct interpretation of the magnitude of a vertical departure of a point from a line depends on where that point appears horizontally—the spread in vertical values expected is roughly proportional to the magnitude of the slope of the assumed quantile function (see Equation (2)). Sometimes this can be ameliorated somewhat by transforming the data first to achieve a flatter quantile function for the test distribution, as happened in the log-normal example transforming from Figure 11(a) to 11(b). The problem is that displaying the points alone provides insufficient information for most users to make good judgments.

By simulating from the test quantile function and adding this information to the plot the above shortcomings dissipate—the qqplot becomes self-calibrating. Envelope bands provide pointwise confidence intervals, exemplar trails allow more detailed shape comparisons, and lineup plots permit a formal visual significance testing.

The approach taken here, to supplement quantile–quantile plots with information based on order statistics simulated from the test distribution, is not new. Birnbaum (1959), for example, had the Applied Mathematics and Statistics Laboratory of Stanford University generated 2500 independent samples from each appropriate half-normal (which changes with the number of experimental contrasts) to estimate critical values that Daniel (1959) would use to serve as "guard rails" on his half-normal plot.

However, the ubiquity of computational resources and display capabilities means that we are no longer constrained to tabulating critical values for each statistic, sample size, and test distribution. The same procedure can be used universally with good effect, particularly for exploratory purposes.

The R package called `qqtest` implements all of the features and examples described above and is freely available to practitioners, instructors, and students everywhere.

[Received December 2014. Revised August 2015.]

REFERENCES

AIAG (1992), *Statistical Process Control* (2nd printing, 1995 ed.), Southfield, MI: Chrysler Corp., Ford Motor Co., and General Motors Corp. [85]

Birnbaum, A. (1959), "On the Analysis of Factorial Experiments Without Replication," *Technometrics*, 1, 343–357. [83,89]

Brownlee, K. A. (1960), *Statistical Theory and Methodology in Science and Engineering*, New York: Wiley. [87]

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., and Wickham, H. (2009), "Statistical Inference for Exploratory Data Analysis and Model Diagnostics," *Philosophical Transactions of the Royal Society,* Series A, 367, 4361–4383. [75,87]

Campbell, N. A. (1980), "Robust Procedures in Multivariate Analysis i: Robust Covariance Estimation," *Journal of the Royal Statistical Society*, Series C, 29, 231–237. [87]

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Boston, MA: Duxbury Press. [81]

Cox, D. R. (1968), "Notes on Some Aspects of Regression Analysis," *Journal of the Royal Statistical Society*, Series A, 131, 265–279. [87]

Cunnane, C. (1978), "Unbiased Plotting Positions—A Review," *Journal of Hydrology*, 37, 205–222. [79]

Daniel, C. (1959), "Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments," *Technometrics*, 1, 311–341. [75,81,82,89]

Daniel, C., and Wood, F. S. (1980), *Fitting Equations to Data: Computer Analysis of Multifactor Data* (2nd ed.), New York: Wiley. [79,81]

Davies, O. L. (ed.) (1956), *The Design and Analysis of Industrial Experiments* (2nd ed.), London: Oliver and Boyd. [82,83]

Dodge, Y. (1996), "The Guinea Pig of Multiple Regression," in *Robust Statistics, Data Analysis, and Computer Intensive Methods Lecture Notes in Statistics* (Vol. 109), ed. H. Rieder, New York: Springer, pp. 591–117 [87]

Fuller, W. E. (1914), "Flood Flows" (with discussion), *Transactions of the American Society of Civil Engineers*, 77, 564–694. [77]

Galton, F. (1875), "Statistics by Intercomparison, With Remarks on the Law of Frequency of Error," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 49, 33–46. [74,75]

——— (1885a), "The Application of a Graphic Method to Fallible Measures," *Journal of the Statistical Society of London,* 262–265. [77]

——— (1885b), "Some Results of the Anthropometric Laboratory," *The Journal of the Anthropological Institute of Great Britain and Ireland*, 14, 275–287. [75]

——— (1885c), "On the Anthropometric Laboratory at the Late International Health Exhibition," *The Journal of the Anthropological Institute of Great Britain and Ireland*, 14, 205–221. [75]

——— (1889), *Natural Inheritance,* London: Macmillan. [74,75]

——— (1899), "A Geometric Determination of the Median Value of a System of Normal Variants, From Two of its Centiles," *Nature*, 61, 102–104. [75,77]

Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations,* New York: Wiley. [81,87]

Gnanadesikan, R., and Kettenring, J. R. (1972), "Robust Estimates, Residuals, and Outlier Detection With Multiresponse Data," *Biometrics*, 28, 81–124. [87]

Hamada, M., and Balakrishnan, N. (1998), "Analyzing Unreplicated Factorial Experiments: A Review With Some New Proposals," *Statistica Sinica*, 8, 1–28. [83]

Hazen, A. (1914), "Storage to be Provided in Impounding Reservoirs for Municipal Water Supply" (with discussion), *Transactions of the American Society of Civil Engineers*, 77, 1539–1669. [74,79,82]

Healy, M. J. R. (1968), "Multivariate Normal Plotting," *Journal of the Royal Statistical Society*, Series C, 17, 157–161. [87]

Lenth, R. V. (2015), "The Case Against Normal Plots of Effects," *Journal of Quality Technology,* 47, 91–97. [83]

Oldford, R. W. (2015), *qqtest: Self Calibrating Quantile-Quantile Plots for Visual Testing,* R Package Version 1.1., available at *https://protect-us.mimecast.com/s/NV68Bwu41pg7ux* [74]

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at *http://www.R-project.org/*. [74]

Rousseeuw, P. J., and Van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633–639. [87]

Smith, T. R. (1884), *Architecture, Gothic and Renaissance*, London: Sampson, Low, Marston, Searle, & Rivington. [75]

Wang, J., Zamar, R., Marazzi, A., Yohai, V., Salibian-Barrera, M., Maronna, R., Zivot, E., Rocke, D., Martin, D., Maechler, M., and Konis, K. (2014), *robust: Robust Library*, R Package Version 0.4–16, available at *https://protect-us.mimecast.com/s/bAlrB1Uk1YOoF6* [79,87]

Whipple, G. C. (1916a), "The Element of Chance in Sanitation," *Journal of the Franklin Institute*, 182, 37–59. [74]

——— (1916b), "The Element of Chance in Sanitation," *Journal of the Franklin Institute*, 182, 205–227. [74,77,79]

Wilk, M. B., and Gnanadesikan, R. (1964), "Graphical Methods for Internal Comparisons in Multiresponse Experiments," *The Annals of Mathematical Statistics*, 35, 613–631. [75,86]

——— (1968), "Probability Plotting Methods for the Analysis of Data," *Biometrika*, 55, 1–17. [81]

Wilk, M. B., Gnanadesikan, R., and Huyett, Miss M. J. (1962), "Probability Plots for the Gamma Distribution," *Technometrics*, 4, 1–20. [87]