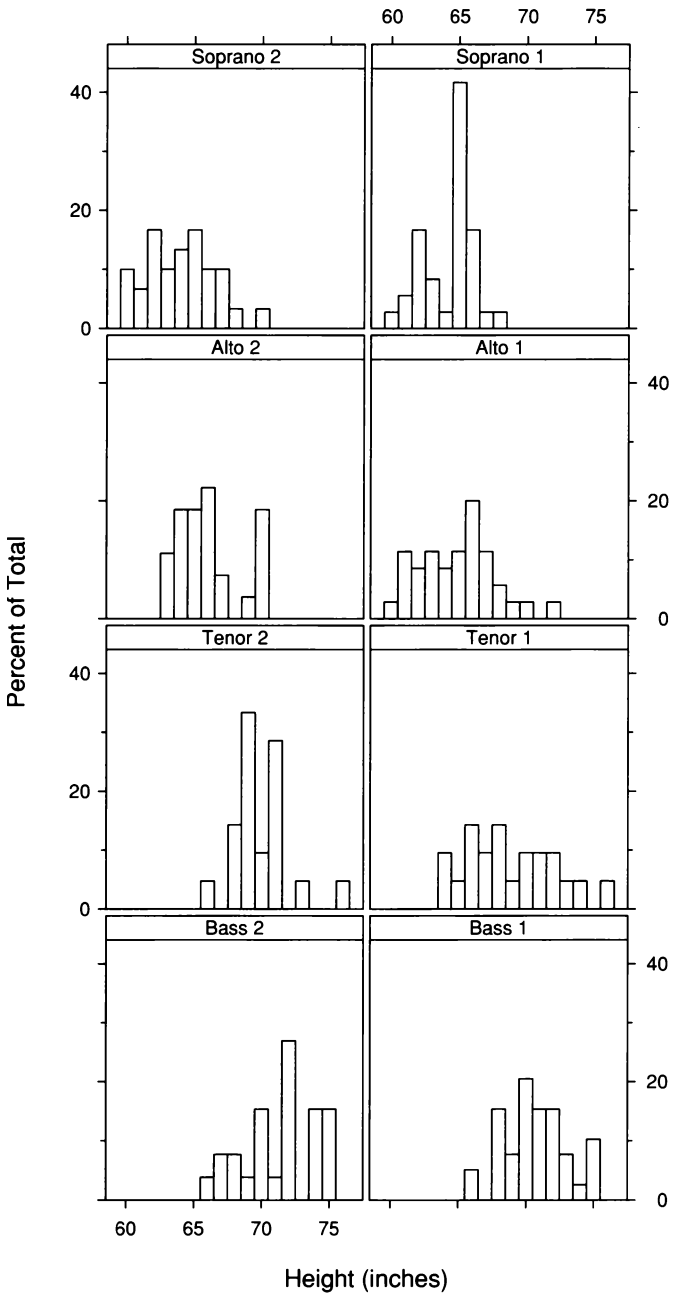
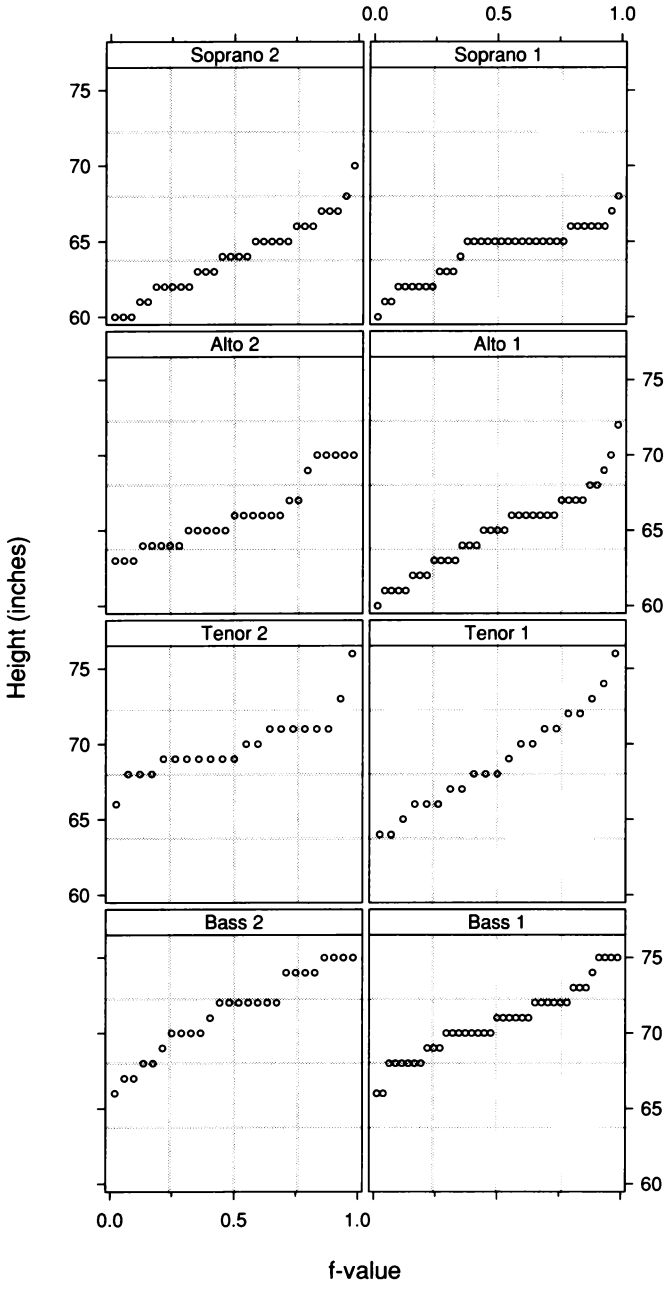


1.2 Histograms graph the singer heights by voice part. The interval width is one inch.



2.1 Quantile plots display univariate data: the heights of singers in the New York Choral Society.

2 *Univariate Data*

Figure 2.1 graphs data introduced in Chapter 1 — heights of singers in the New York Choral Society [16]. The vertical scale on the panels is height; the horizontal scale, which will be explained shortly, ranges from 0 to 1. The singers are divided into eight voice parts, shown by the panel labels in the figure. The altos and sopranos are women and the tenors and basses are men. The pitch intervals of the voice parts are different. Starting from the lower left panel in Figure 2.1, the pitch interval increases as we progress through the panels from left to right and from bottom to top. For example, the second basses must sing lower pitches than the first basses, who in turn sing lower pitches than the second tenors.

The singer heights are *univariate data*: measurements of a single quantitative variable. The measurements are broken up into groups by the categorical variable, voice part. The goal in analyzing the singer data is to determine whether voice part is related to height. One might expect this because taller people tend to be larger overall, and larger vocal tracts would have lower resonance frequencies, and thus produce lower tones.

2.1 *Quantile Plots*

The singer heights for each voice part occupy positions along the measurement scale. The collection of positions is the *distribution* of the data. Thus the goal in analyzing the data is to compare the eight height distributions.

Quantiles

Quantiles are essential to visualizing distributions. The f quantile, $q(f)$, of a set of data is a value along the measurement scale of the data with the property that approximately a fraction f of the data are less

than or equal to $q(f)$. The property has to be approximate because there might not be a value with exactly a fraction f of the data less than or equal to it. The 0.25 quantile is the *lower quartile*, the 0.5 quantile is the *median*, and the 0.75 quantile is the *upper quartile*.

The graphical methods in this chapter are largely visualizations of quantile information. Quantiles provide a powerful mechanism for comparing distributions because f-values provide a standard for comparison. To compare distributions we can compare quantiles with the same f-values. For example, for the singer heights, the median of the second basses is 72 inches, 4 inches greater than the median of the first tenors. This is a meaningful and informative comparison.

An explicit rule is needed for computing $q(f)$. Consider the first tenor heights. Let $x_{(i)}$, for $i = 1$ to n, be the data ordered from smallest to largest; thus $x_{(1)}$ is the smallest observation and $x_{(n)}$ is the largest. For the first tenors, $n = 21$, $x_{(1)} = 64$ inches, and $x_{(21)} = 76$ inches. Let

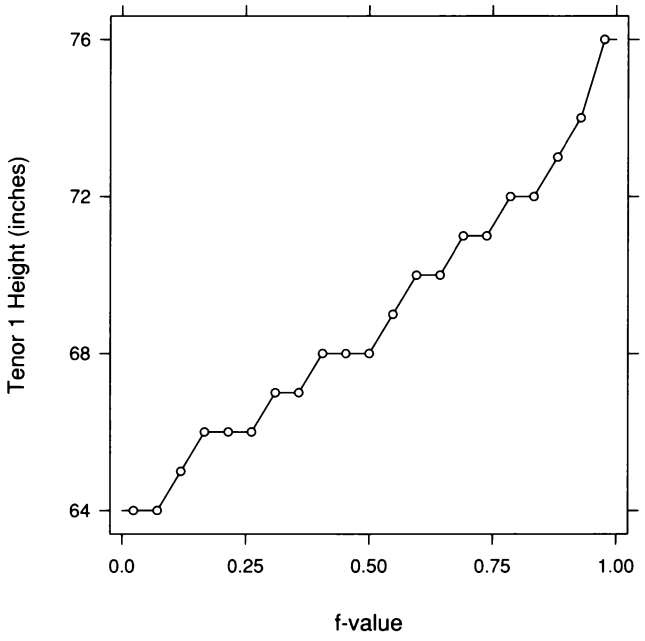
$$f_i = \frac{i - 0.5}{n} .$$

These numbers increase in equal steps of $1/n$ beginning with $1/2n$, which is slightly above zero, and ending with $1 - 1/2n$, which is slightly below one. For the first tenors, the values go from $1/42$ to $41/42$ in steps of $1/21$. We will take $x_{(i)}$ to be $q(f_i)$. For the first tenors the values are

f	x	f	x	f	x
0.02	64	0.36	67	0.69	71
0.07	64	0.40	68	0.74	71
0.12	65	0.45	68	0.79	72
0.17	66	0.50	68	0.83	72
0.21	66	0.55	69	0.88	73
0.26	66	0.60	70	0.93	74
0.31	67	0.64	70	0.98	76

The precise form of f_i is not important; we could have used $i/(n+1)$, or even i/n, although this last value would prove inconvenient later for visualization methods that employ the quantiles of a normal distribution.

So far, $q(f)$ has been defined just for f-values that are equal to f_i. The definition is extended to all values of f from 0 to 1 by linear interpolation and extrapolation based on the values of f_i and $q(f_i)$. Figure 2.2 illustrates the method using the first tenor heights. The plotting symbols are the points $(f_i, x_{(i)})$; the interpolation and extrapolation are shown by the line segments.



2.2 The symbols and the line segments show the quantile function of the first tenor heights.

Graphing Quantiles

On a *quantile plot*, $x_{(i)}$ is graphed against f_i. In other words, we visualize the f_i quantiles. The panels of Figure 2.1 are quantile plots of the singer heights. The interpolated and extrapolated values of $q(f)$ are not shown because they do not appreciably enhance our visual assessment of the distribution of the data. Rather, the interpolation or extrapolation is used when, for some other purpose, we need a quantile whose f-value does not happen to be one of the values of f_i.

Graphing univariate measurements on a quantile plot is a simple and effective way to have a first look at their distribution. First, the values of all of the data are displayed; we can assess both overall behavior and unusual occurrences. And information about quantiles is conveyed.

Figure 2.1 shows several properties of the singer data: the heights are rounded to the nearest inch; the values for each voice part have a reasonably wide range, about one foot in some cases; there are many first soprano heights equal to 65 inches; and the data have no particularly unusual values, for example, no exceptionally tall or short singers. The height distributions vary substantially with the voice part. At one extreme, the median height is 65 inches for first sopranos, diminutive women piercing the air with notes as high as two octaves above middle C. At the other extreme, the median height is 72 inches for the second basses, tall men vibrating the stage with notes as low as two octaves below middle C. Shortly, other methods for visualizing quantiles will reveal more about the shift in the distributions.

Graphical Order and Visual Reference Grids

Figure 2.1 uses an important convention that will be followed in the remainder of the book; when the panels of a multi-panel display are associated with an ordered variable, such as pitch interval, the variable will increase as we go from left to right and from bottom to top. If the ordered variable were graphed in some way along a horizontal scale, it would increase in going from left to right; if the variable were graphed in some way along a vertical scale, it would increase in going from bottom to top. The *graphical order* of the panels simply follows the established convention.

Figure 2.1 has *visual reference grids*, the vertical and horizontal lines in gray. Their purpose is not to enhance scale reading, or *table look-up*, which is the determination of numerical values from the scales; the tick marks are sufficient for table look-up. Rather, their purpose is to enhance the comparison of patterns, or gestalts, on different panels. By providing a common visual reference, the grids enhance our comparison of the relative locations of features on different panels [21]. For example, in Figure 2.1, the grids make it easy to see that almost all second basses are taller than all of the first sopranos.

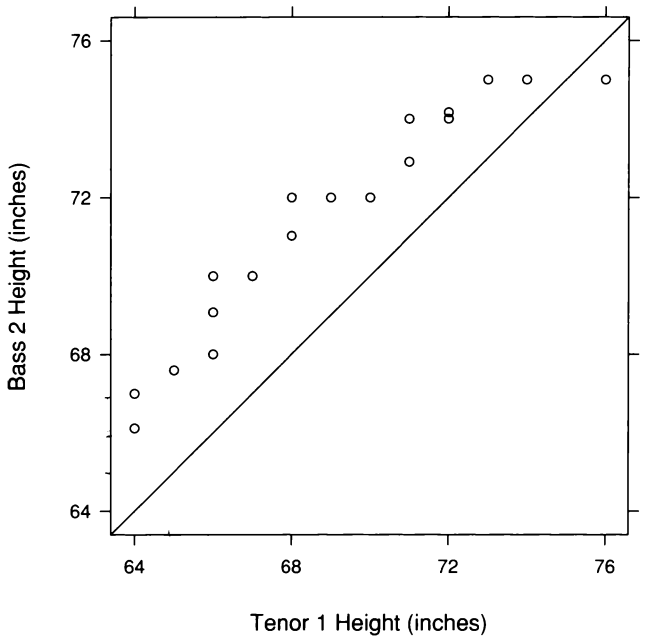
2.2 Q-Q Plots

The *quantile-quantile plot*, or *q-q plot*, of Wilk and Gnanadesikan is a powerful visualization method for comparing the distributions of two or more sets of univariate measurements [80]. When distributions are compared, the goal is to understand how the distributions *shift* in going from one data set to the next. For the singers, the goal is to understand how the height distributions shift with voice part.

The most effective way to investigate the shifts of distributions is to compare corresponding quantiles. This was the insightful observation of Wilk and Gnanadesikan, and their invention could not be more simple or elegant — two distributions are compared by graphing quantiles of one distribution against the corresponding quantiles of the other.

Suppose there are just two sets of univariate measurements to be compared. Let $x_{(1)}, \ldots, x_{(n)}$ be the first data set, ordered from smallest to largest. Let $y_{(1)}, \ldots, y_{(m)}$ be the second, also ordered. Suppose $m \leq n$. If $m = n$, then y_i and x_i are both $(i - 0.5)/n$ quantiles of their respective data sets, so on the q-q plot, $y_{(i)}$ is graphed against $x_{(i)}$; that is, the ordered values for one set of data are graphed against the ordered values of the other set. If $m < n$, then y_i is the $(i - 0.5)/m$ quantile of the y data, and we graph y_i against the $(i - 0.5)/m$ quantile of the x data, which typically must be computed by interpolation. With this method, there are always m points on the graph, the number of values in the smaller of the two data sets. Of course, if m is a big number say 10^3, then we can select fewer quantiles for comparison.

Figure 2.3 graphs quantiles of the 26 second basses against quantiles of the 21 first tenors. The size of the smaller data set is 21, so 21 quantiles with f-values equal to $(i - 0.5)/21$ are compared. Because some of the plotting symbols overlap, only 18 distinct points appear on the graph. Ordinarily, such overlap would require a remedy discussed in Chapter 3 — jittering, an addition of uniform random noise to coordinates of the points. But on a q-q plot, the points portray what is actually an increasing continuous curve, one quantile function against another, so breaking up the overlap is not necessary. When $i = 1$, the f-value is 0.024. The point in the lower left corner of the data region is the 0.024 quantile for the basses against the 0.024 quantile for the tenors. When $i = 21$, the f-value is 0.976. The point in the upper right corner of the data region is the 0.976 quantile for the basses against the 0.976 quantile for the tenors.



2.3 The first tenor and second bass height distributions are compared by a q-q plot.

The line in Figure 2.3 is $b = t$, where b stands for bass and t stands for tenor. Let $(t_i,\ b_i)$ be the coordinates of the points graphed on the panel. Our goal in studying the q-q plot is to determine how the points deviate from the line $b = t$. If the distributions of the tenor and bass heights were the same, the points would vary about this line. But they are not the same. There is a shift between the distributions; the underlying pattern of the points is a line

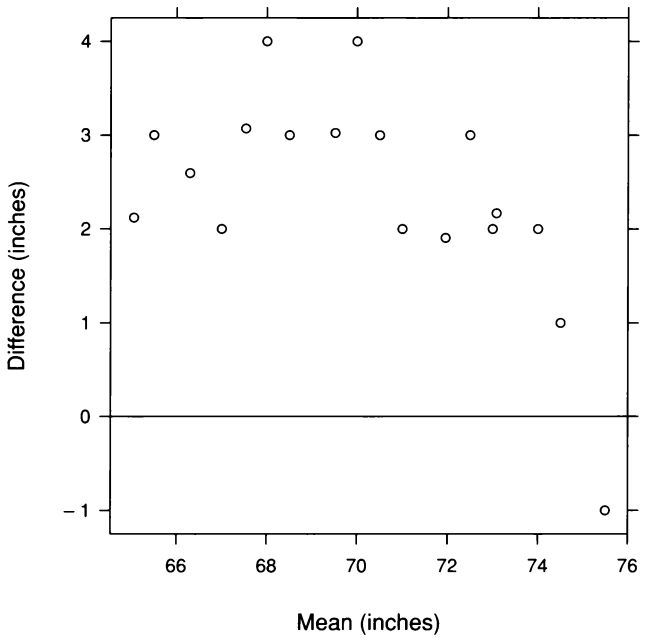
$$b = t + c.$$

Before interpreting this pattern, we will make one more graph to visualize it in a different way.

Tukey Mean-Difference Plots

A *Tukey mean-difference plot*, or *m-d plot*, can add substantially to our visual assessment of a shift between two distributions. Figure 2.4 is an m-d plot derived from the q-q plot in Figure 2.3. The differences, $b_i - t_i$,

are graphed against the means, $(b_i + t_i)/2$. The line $b = t$ on the q-q plot becomes the zero line on the m-d plot, and a shift is assessed by judging deviations from the zero line. This often enhances our perception of effects because we can more readily judge deviations from a horizontal line than from a line with nonzero slope.

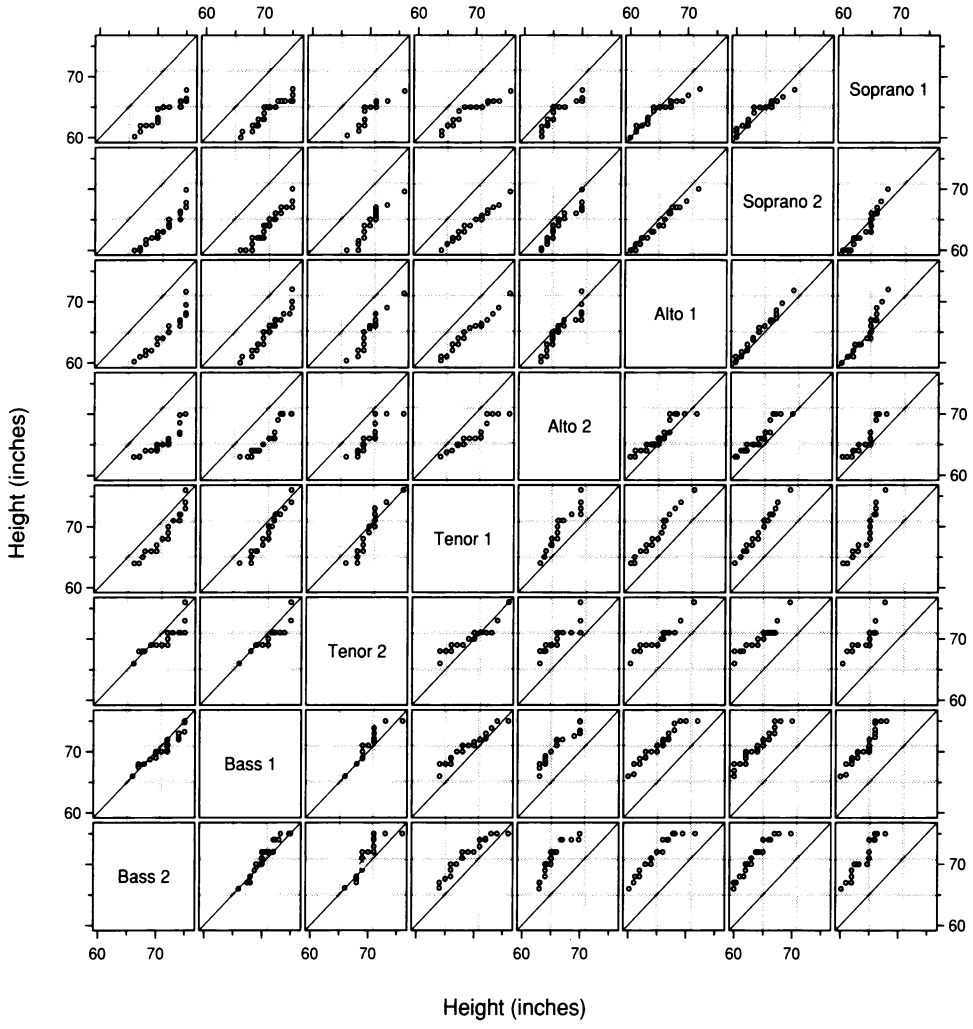


2.4 The first tenor and second bass height distributions are compared by a Tukey m-d plot.

Figures 2.3 and 2.4 show that the tenor and bass distributions differ in an exceedingly simple way: the quantiles of the bass distribution are roughly equal to the quantiles of the tenor distribution plus a constant of about 2.5 inches. There is an *additive shift* of about 2.5 inches. The comparison of the two distributions can be summarized by the simple statement that the distribution of the bass heights is about 2.5 inches greater. This is good news; later examples will show that shifts between distributions can be complex.

Pairwise Q-Q Plots

The goal in analyzing the singer data is to compare the distributions of all voice parts and determine the shifts. Figure 2.5 shows the q-q plots of all possible pairs of voice parts. For example, the second row from the bottom has q-q plots of the first basses against all other voice parts. The second column from the left also compares the distribution of the first basses with all others, but now the first bass quantiles are on the horizontal axis instead of the vertical axis. Thus we can scan either the second row or column to compare the first basses with all other voice parts.

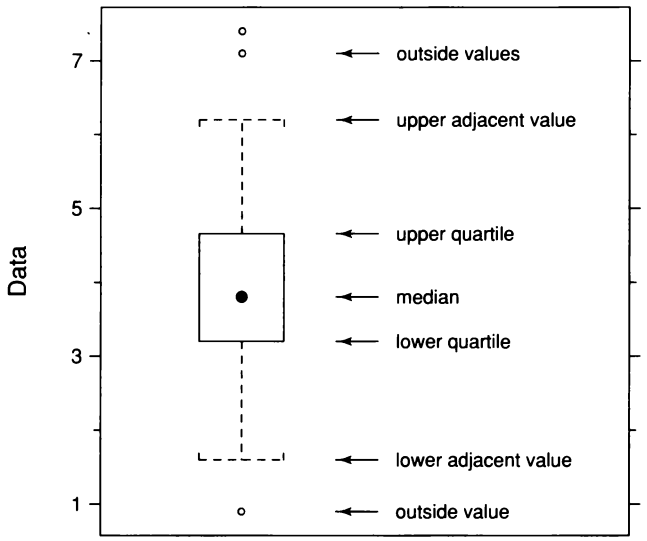


2.5 The height distributions are compared by q-q plots for all pairs of singer voice parts.

There is a great deal of information in Figure 2.5. Overall, there is a suggestion that shifts between pairs of distributions are additive. But we need a way to distill the information because there are so many pairs to compare. Coming methods help with the distillation.

2.3 *Box Plots*

One method for distilling the information on q-q plots is Tukey's *box plot* [76]. Instead of comparing many quantiles, as on the q-q plot, a limited number of quantities are used to summarize each distribution, and these summaries are compared.

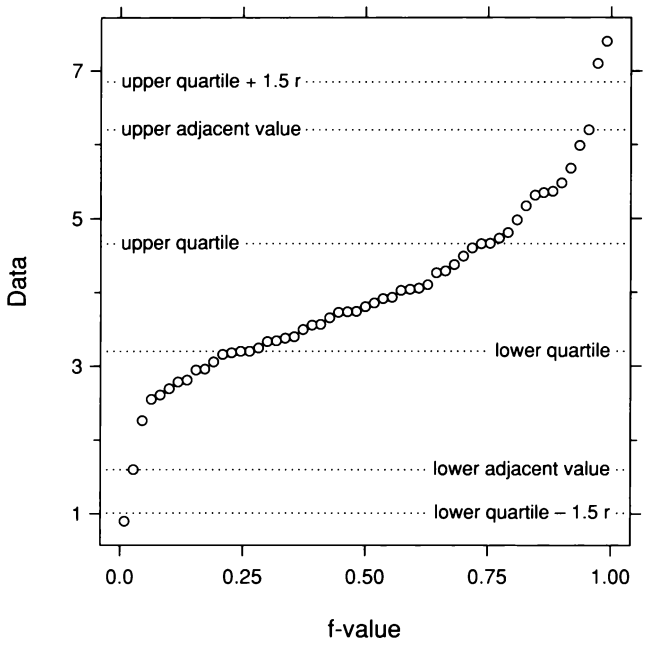


2.6 The diagram defines the box plot display method.

The method of summary is illustrated in Figure 2.6. The filled circle encodes the median, a measure of the center, or *location*, of the distribution. The upper and lower ends of the box are the upper and lower quartiles. The distance between these two values, which is the *interquartile range*, is a measure of the *spread* of the distribution. The middle 50% or so of the data lie between the lower and upper quartiles. If the interquartile range is small, the middle data are tightly packed around the median. If the interquartile range is large, the middle data spread out far from the median. The relative distances of the upper

and lower quartiles from the median give information about the *shape* of the distribution of the data. If one distance is much bigger than the other, the distribution is *skewed.*

The dashed appendages of the box plot encode the *adjacent values.* Let r be the interquartile range. The upper adjacent value is the largest observation that is less than or equal to the upper quartile plus $1.5r$. The lower adjacent value is the smallest observation that is greater than or equal to the lower quartile minus $1.5r$. Figure 2.7, a quantile plot, demonstrates their computation. The adjacent values also provide summaries of spread and shape, but do so further in the extremes, or *tails,* of the distribution.

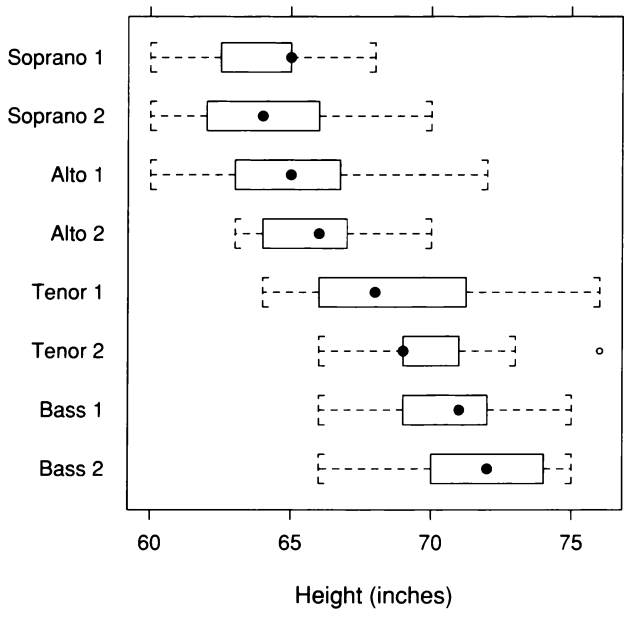


2.7 The diagram illustrates the computation of the adjacent values, which are used in the box plot display method.

Outside values, observations beyond the adjacent values, are graphed individually. Sometimes, the upper adjacent value is the maximum of the data, so there are no outside values in the upper tail; a similar statement holds for the lower tail. Outside values portray behavior in the extreme tails of the distribution, providing further information about

spread and shape. If there happen to be outliers — unusually large or small observations — they appear as outside values, so the box plot method of summarizing a distribution does not sweep outliers under the rug.

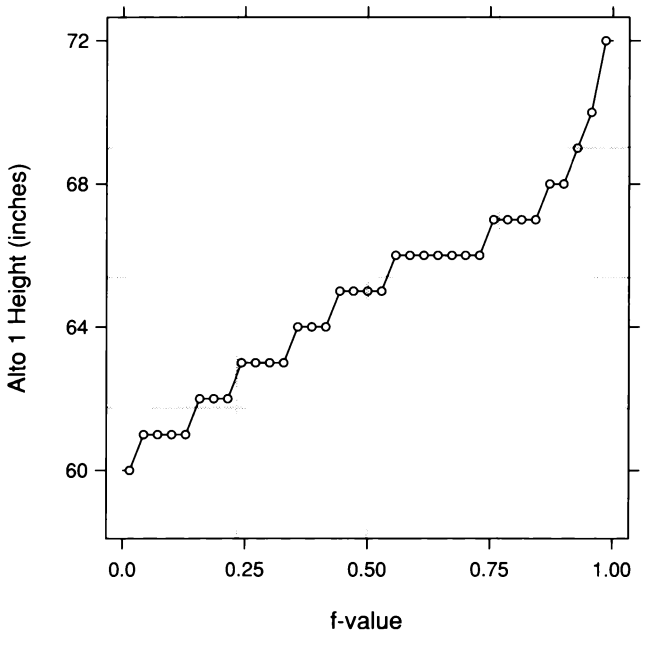
Figure 2.8 shows box plots of the singer heights. The visualization effectively conveys the relationship of voice part and height. Part of the reason is the perceptual effectiveness of the display method. Each specific aspect of the distributions that is encoded — medians, upper quartiles, lower quartiles, and so forth — can be readily visually decoded [21, 77]. For example, it is easy to see the medians as a whole, and visually assess the values. Also, the format is horizontal to enhance the readability of the labels. Figure 2.8 shows that for the men singers, the tenors and basses, height tends to decrease as the pitch interval increases; the same is true of the women singers, the sopranos and altos. One exception to this pattern is the first sopranos, whose height distribution is quite similar to that of the second sopranos. This might be due to a bias in the measurements; the collector of the data noticed a tendency for the shortest sopranos to round their heights strongly upward.



2.8 The eight singer distributions are compared by box plots.

2.4 Normal Q-Q Plots

The box plots in Figure 2.8 show certain effects that accord with our prior knowledge of height and pitch interval, and that we would expect to be reproduced if we were to study another set of singer heights. One such reproducible effect is the general decrease in the height distributions with increasing pitch interval. But the box plots also show spurious effects: unreproducible variation, and variation that is an artifact of rounding to the nearest inch. This spurious variation does not speak to the relationship of height and voice part. For example, for the first sopranos, the median is equal to the upper quartile; for the second tenors, the median is equal to the lower quartile; and the two tallest singers are a first tenor and a second tenor. There is nothing to be learned about height and voice part from these properties of the data. In this section and the next, we move from the graphical methods of the previous sections, which give full vent to the variation in the data, both informative and spurious, to methods whose purpose is to impose structure on the data in an attempt to help us decide what variation appears meaningful and what variation it is better to ignore.

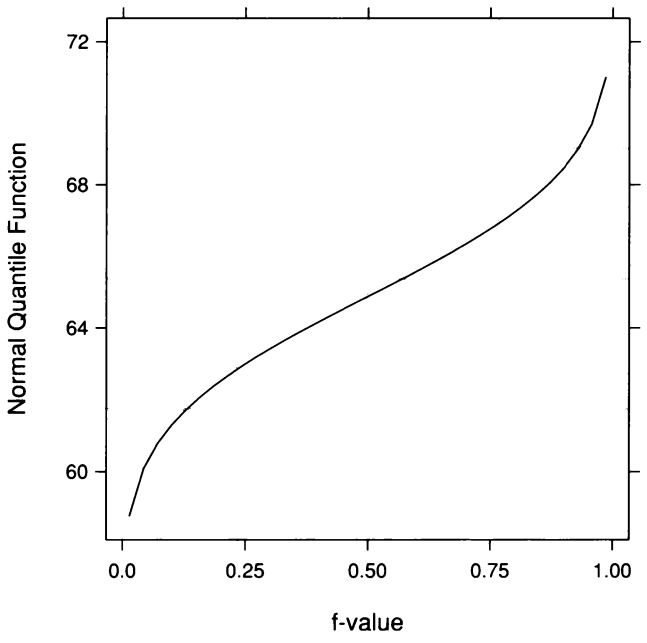


2.9 The plotting symbols and the line segments display the quantile function of the first altos.

Data Quantiles and Normal Quantiles

Figure 2.9 graphs the quantile function, $q(f)$, of the first alto heights. This is the quantile function of a data distribution, a set of real-life univariate measurements. Figure 2.10 graphs quantiles of a mathematical distribution, the normal distribution, a staple of probabilistic inference. It has no reality in the sense that no set of data could ever have such a normal distribution or even be construed as genuinely being a sample from a normal population of potential measurements. It implies, among other things, that measurements range from $-\infty$ to ∞ and have infinite accuracy. Yet it is still helpful to check if the normal quantile function, however imaginary, serves as a reasonable approximation to the real thing, the data distribution.

The normal distribution is a family of distributions. Each normal distribution in the family is specified by two numerical values: the mean, μ, and the standard deviation, σ. The mean is a measure of the location of the distribution, and the standard deviation is a measure of the spread of the distribution about its mean. Given μ and σ, we can



2.10 The curve displays the quantiles of a normal distribution for f-values from 1/70 to 69/70, which are the f-values of the minimum and maximum heights of the first altos.

compute the quantile function, $q_{\mu,\sigma}(f)$, of the specified distribution; in Figure 2.10, the specified mean is the sample mean of the first alto heights,

$$\bar{x} = \frac{1}{35}\sum_{i=1}^{35} x_i = 64.9 \text{ inches},$$

and the specified standard deviation is the sample standard deviation of these heights,

$$s = \sqrt{\frac{1}{34}\sum_{i=1}^{35}(x_i - \bar{x})^2} = 2.8 \text{ inches}.$$

Most people think about the normal distribution in terms of random variables and probabilities. Suppose nature generates a value, x, of a random variable with a normal distribution. Let f be a probability between 0 and 1. Then the probability that x is less than or equal to $q_{\mu,\sigma}(f)$ is f. That is, a fraction f of the mass of the normal distribution is less than or equal to $q_{\mu,\sigma}(f)$. Notice that the definition of the normal quantile is analogous to the definition of the quantile, $q(f)$, of a set of data; approximately a fraction f of the mass of the data is less than or equal to $q(f)$.

Graphing Data Quantiles and Normal Quantiles

A *normal quantile-quantile plot*, or *normal q-q plot*, is a graphical method for studying how well the distribution of a set of univariate measurements is approximated by the normal. As before, let $x_{(i)}$ be the data, ordered from smallest to largest, and let $f_i = (i - 0.5)/n$. Suppose the distribution of the data is well approximated by some normal distribution with mean μ and standard deviation σ. $x_{(i)}$ is the f_i

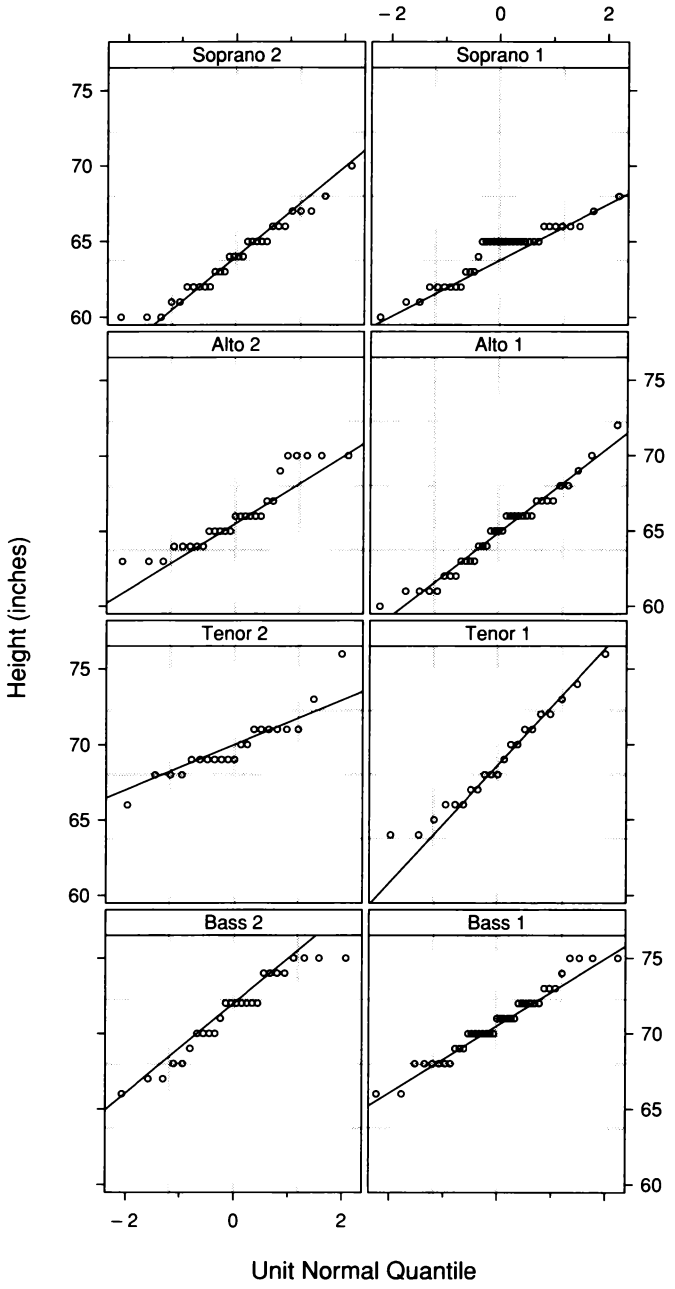
quantile of the data, and $q_{\mu,\sigma}(f_i)$ is the corresponding normal f_i quantile of the approximating distribution. If $x_{(i)}$ is graphed against $q_{\mu,\sigma}(f_i)$, the overall pattern is a line with intercept 0 and slope 1. An important property of normal quantiles is that

$$q_{\mu,\sigma}(f) = \mu + \sigma q_{0,1}(f) \, ,$$

where $q_{0,1}(f)$ is the quantile function of the unit normal distribution, which has $\mu = 0$ and $\sigma = 1$. In other words, to go from unit normal quantiles to general normal quantiles, we simply multiply by σ and add μ. Thus, if the distribution of the $x_{(i)}$ is well approximated by a normal distribution, then on a graph of $x_{(i)}$ against $q_{0,1}(f_i)$, the overall pattern is a line with intercept μ and slope σ.

A graph of $x_{(i)}$ against $q_{0,1}(f_i)$ is a normal q-q plot. To judge the normal approximation, we judge the underlying pattern of the points on the graph. If the pattern is linear, or nearly so, the data distribution is well approximated by the normal. If not, the deviations from linearity convey important information about how the data distribution deviates from the normal.

Figure 2.11 shows normal q-q plots of the singer data. On each panel, a line is superposed to help us judge the straightness of the pattern. The line passes through the upper and lower quartiles, $(q_{0,1}(0.25), q(0.25))$ and $(q_{0,1}(0.75), q(0.75))$. In each case, the overall pattern appears nearly straight; that is, the eight distributions are reasonably well approximated by normal distributions. The deviations from the overall pattern are inflated by rounding to the nearest inch, which produces strings of points positioned at integer values along the vertical scale. Such discreteness is a departure from normality, because the normal distribution is a continuous one in which any value is possible. But the rounding in this case is not so severe that the approximation is seriously jeopardized.



2.11 Normal q-q plots compare the eight height distributions with the normal distribution.

The Normal Approximation

Many good things happen when data distributions are well approximated by the normal. First, the question of whether the shifts among the distributions are additive becomes the question of whether the distributions have the same standard deviation; if so, the shifts are additive. The slope of the pattern of points on a normal q-q plot is an indicator of the standard deviation of the approximating normal, so judging whether standard deviations are equal from normal q-q plots involves judging whether the slopes are equal. For the singer data, the slopes in Figure 2.11 do vary, but not by a large amount. Most importantly, the variation in the slopes is not related to the means of the distributions. If there were a meaningful change in the standard deviations of the singer distributions, we would expect it to take the form of an increase in the voice-part standard deviations as the voice-part means increase. This is not the case.

A second good happening is that methods of fitting and methods of probabilistic inference, to be taken up shortly, are typically simple and on well understood ground. For example, statistical theory tells us that the sample mean, $\bar{x}$, provides a good estimate of the location of the distribution, and the sample standard deviation, s, provides a good estimate of the spread of the distribution.

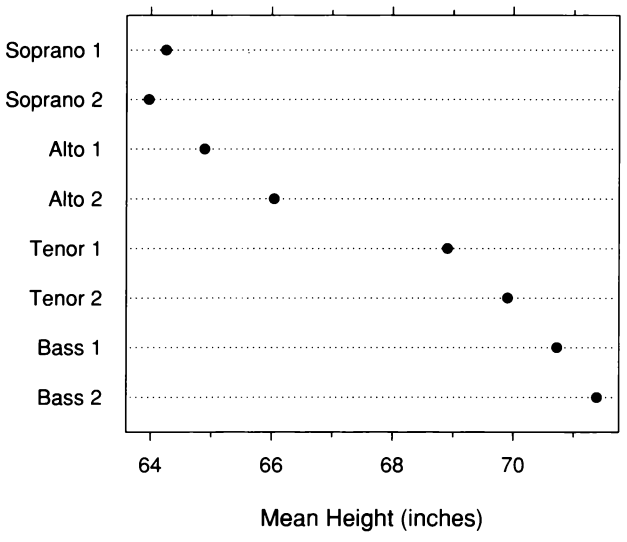
A third good thing is that the description of the data distribution is more parsimonious. A data distribution needs n values to completely determine it, the n observations x_i. If the distribution is well approximated by the normal, these n values can be replaced by two values, $\bar{x}$ and s. The quantiles of the data are then described by the quantiles of the normal distribution with a mean equal to $\bar{x}$ and a standard deviation equal to s. For example, with these values, we know that the upper quartile of the data is about $\bar{x} + 0.67s$.

2.5 *Fits and Residuals*

Fitting data means finding mathematical descriptions of structure in the data. An additive shift is a structural property of univariate data in which distributions differ only in location and not in spread or shape. For example, the visualization of the singer data has so far suggested that the voice-part distributions differ only in location. An additive shift is fitted by *estimating* location — computing a location measure for each distribution.

Fitting Additive Shifts by Location Estimates

Two candidates for location estimation are the median and the mean. For the singer data, since the normal q-q plots showed that the distributions are well approximated by the normal, we will follow the imperatives of statistical theory and use means. The voice-part means are graphed by a dot plot in Figure 2.12. The mean height decreases with increasing pitch interval, except for the means of the first and second sopranos, which are very nearly equal. We saw this property earlier in the box plots of Figure 2.8, which used medians rather than means as the location measure. At that stage of our analysis, which was preliminary and exploratory, we used medians because they are not distorted by a few outliers, a property not shared by the mean; we will return to this issue of distortion in Section 2.8.



2.12 A dot plot displays the sample means of the height distributions.

Fitted Values and Residuals

For the pth voice part of the singer data, let h_{pi} be the ith measurement of height and let $\bar{h}_p$ be the mean height. In fitting the eight voice-part means to describe the additive shifts, each singer height has had its voice-part mean fitted to it. The *fitted value* for h_{pi} is

$$\widehat{h}_{pi} = \bar{h}_p.$$

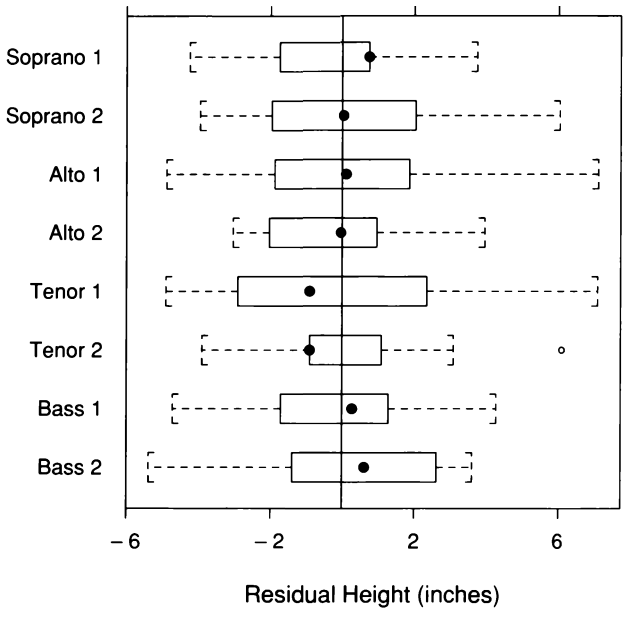
The *residuals* are the deviations of the heights from the fitted values,

$$\widehat{\varepsilon}_{pi} = h_{pi} - \widehat{h}_{pi}.$$

Thus the heights have been decomposed into two parts,

$$h_{pi} = \widehat{h}_{pi} + \widehat{\varepsilon}_{pi}.$$

The fitted values account for the variation in the heights attributable to the voice-part variable through the fitting process. The residuals are the remaining variation in the data after the variation due to the shifting means has been removed. This removal is shown in Figure 2.13, which graphs the eight residual distributions. Since the subtraction of means has removed the effect of location, the box plots are centered near zero.



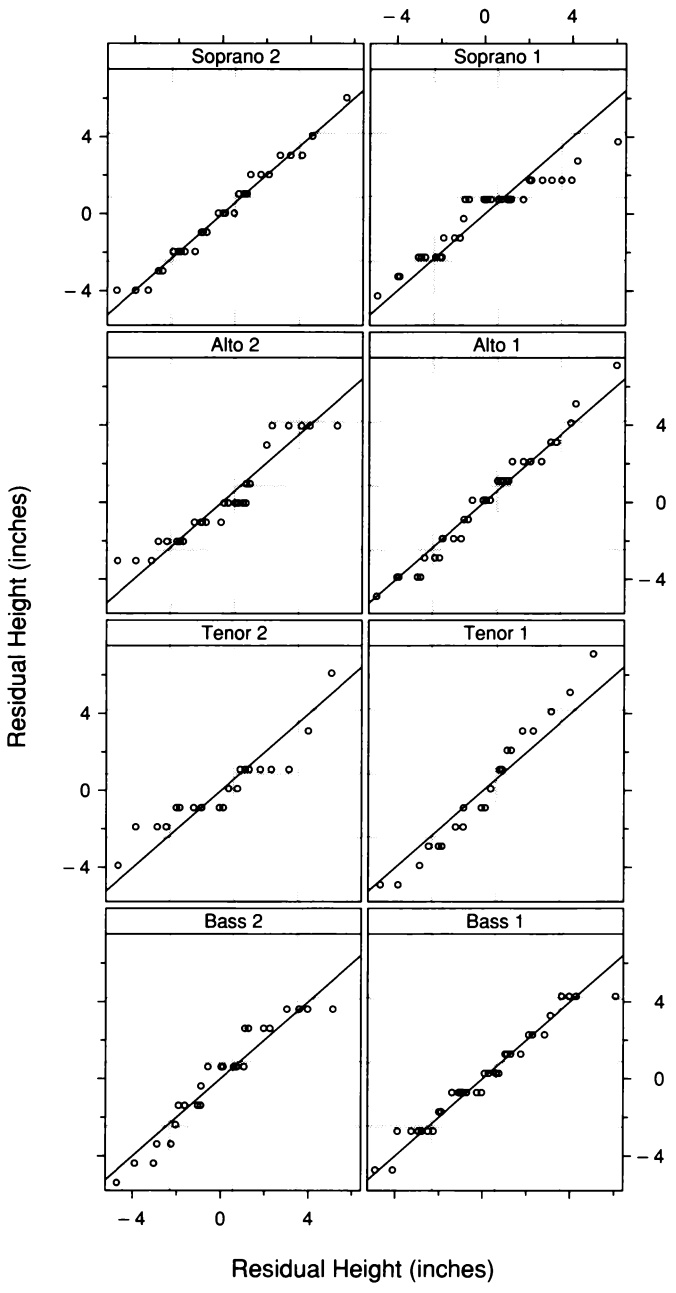
2.13 Box plots compare the distributions of the height residuals for the fit to the data by voice-part means.

Suppose, as our visualization so far suggests, that the underlying patterns of the eight singer distributions differ by additive shifts. Then the distributions of the eight sets of residuals should be nearly the same because subtracting means eliminates the shifts. Our next task is to compare the distributions of the residuals to see if they appear to be nearly the same; this provides a confirmation of our additive-shift observation.

The residual distributions could be compared by all pairwise q-q plots, but there are 28 pairs, and we would be back to the problem of the pairwise q-q plots of the data in Figure 2.5 — assessing a substantial amount of variation. Figure 2.14 uses another method that results in just eight q-q plots. On each panel, the quantiles of the residuals for one voice part are graphed on the vertical scale against the quantiles of the residuals for all voice parts on the horizontal scale. The line on each panel has slope one and intercept zero. Since the underlying patterns of the points on the eight panels follow these lines, the residual distributions are about the same. This adds credence to a conclusion of additive shifts.

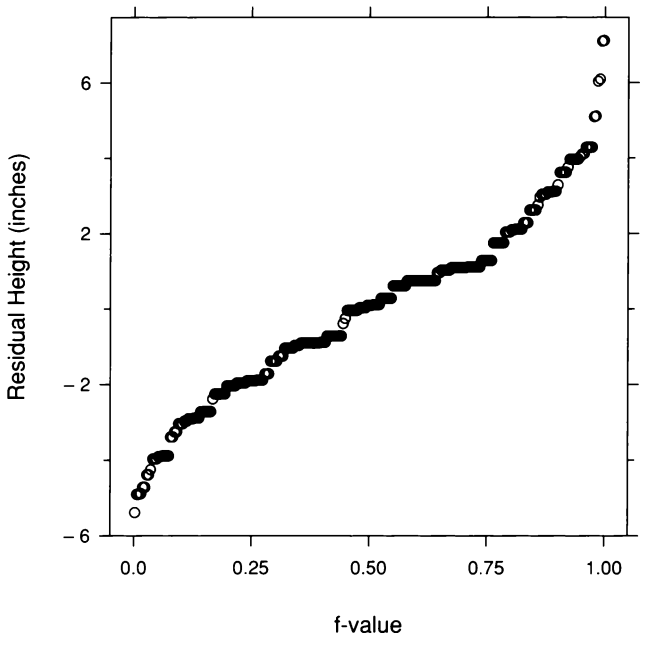
Homogeneity and Pooling

The process of identifying a structure in data and then fitting the structure to produce residuals that have the same distribution lies at the heart of statistical analysis [3, 30, 35, 76]. Such *homogeneous* residuals can be *pooled*, which increases the power of the description of the variation in the data. For the singer data, we have judged the eight residual distributions to be homogeneous, so we can pool them, using the variation in the entire collection of residuals to describe the variation in the residuals for each voice part. This fitting and pooling leads, as we will now show, to a more informative characterization of the variation of the height distribution for each of the voice parts.



2.14 Each panel is a q-q plot that compares the distribution of the residuals for one voice part with the distribution of the pooled residuals.

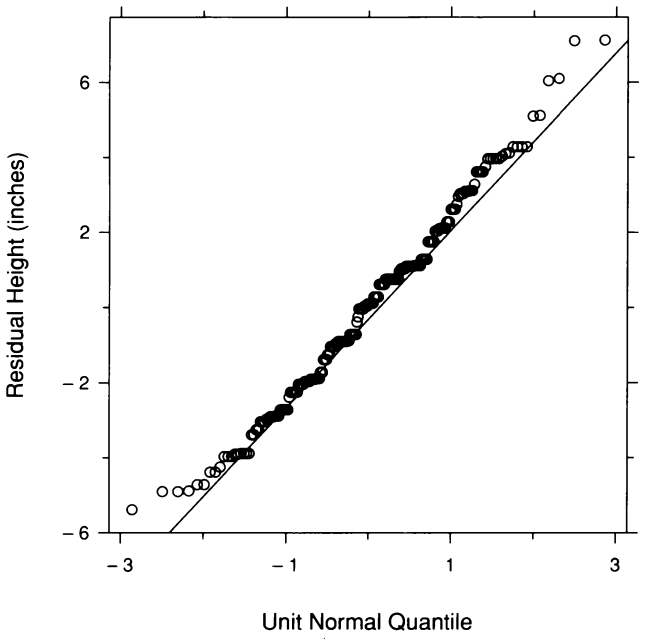
Each voice-part distribution can be described by its mean together with the pooled residual variation. Figure 2.15 is a quantile plot of all of the residuals. The 0.025 quantile of the residuals is -4.6 inches and the 0.975 quantile is 4.8 inches. Thus 95% of the residuals lie between about ± 4.7 inches. Consider the first sopranos. The mean is 64.3 inches. The resulting description of the 95% first soprano variation is 64.3 ± 4.7 inches, which is 59.6 inches to 69.0 inches. Similarly, the second bass variation is 71.4 ± 4.7 inches, which is 66.7 inches to 76.1 inches. We have been able to use the richer information source of all of the residuals to describe the variation in the second basses and the variation in the first sopranos, rather than relying on just the first soprano heights to describe their variation, and just the second bass heights to describe their variation. Of course, the pooling power has come from imposing structure on the data — an additive-shift structure — and the resulting description is valid only if the structure is valid, but the visualization of the data has made the imposition entirely reasonable.



2.15 The pooled residuals are displayed by a quantile plot.

Fitting the Normal Distribution

Once the homogeneity of a set of residuals has been established, we can attempt a fit of the normal distribution to them. Figure 2.16 is a normal quantile graph of the singer residuals. There is a hint of curvature in the underlying pattern, but the effect is relatively minor, so the residual distribution is reasonably well approximated by the normal. Thus we can use a fitted normal to characterize the variability in the residuals.



2.16 A normal q-q plot compares the distribution of the pooled residuals with a normal distribution.

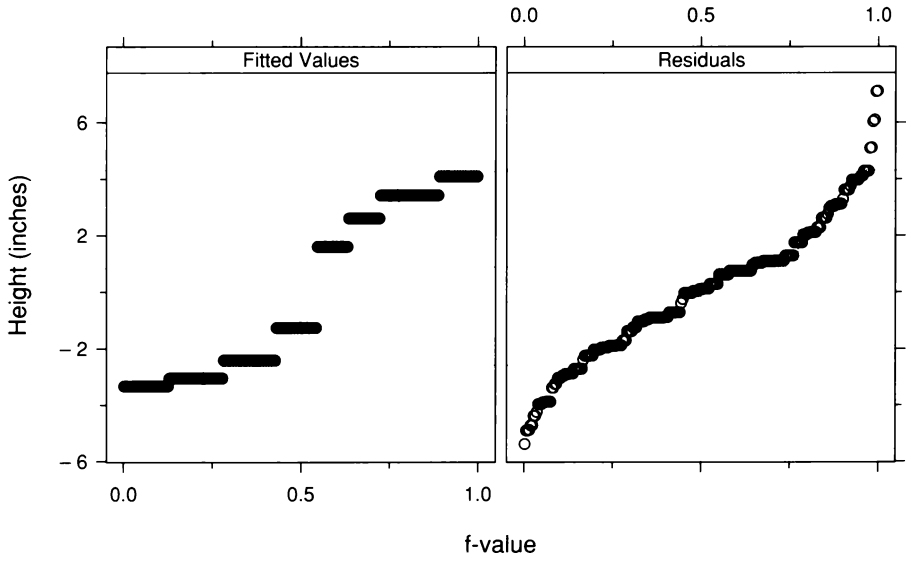
The sample mean of the residuals is 0 inches because the mean of the residuals for each voice part is 0 inches. The sample standard deviation of the residuals is

$$\sqrt{\frac{1}{n-8}\sum_{p=1}^{8}\sum_{i}\hat{\varepsilon}_{pi}^{2}} = 2.45 \text{ inches}.$$

Thus the fitted normal has a mean of 0 inches and has a standard deviation of 2.45 inches. The 95% variation about the mean of a normal distribution is ± 1.96 times the standard deviation; for the singer residuals, this is ± 4.8 inches, a value that is very close to the ± 4.7 inches that arose from the residual quantiles. Thus, using this normal fit, the description of the 95% variability in the first soprano heights is 59.5 inches to 69.1 inches and the 95% variability in the second bass heights is 66.6 inches to 76.2 inches. The description via the normal approximation is attractive because it is more parsimonious than the description based on the residual quantiles.

The Spreads of Fitted Values and Residuals

It is informative to study how influential the voice-part variable is in explaining the variation in the height measurements. The fitted values and the residuals are two sets of values each of which has a distribution. If the spread of the fitted-value distribution is large compared with the spread of the residual distribution, then the voice-part variable is influential. If it is small, the voice-part variable is not as influential. Figure 2.17 graphs the two distributions by quantile plots; since it is the spreads of the distributions that are of interest, the fitted values minus their overall mean are graphed instead of the fitted values themselves. This *residual-fit spread plot*, or *r-f spread plot*, shows that the spreads of the residuals and the fitted values are comparable. Thus the voice-part variable accounts for a significant portion of the variation in the height data, but there is a comparable amount of variation remaining.



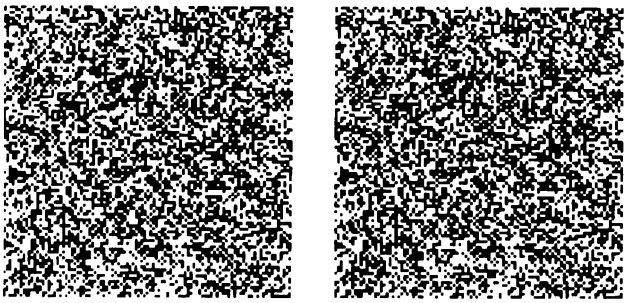
2.17 An r-f spread plot compares the spreads of the residuals and the fitted values minus their mean for the fit to the singer data.

Fitting and Graphing

This approach to studying the singer distributions — fitting sample means to the data, subtracting them to get residuals, and then graphing the fit and residuals — illustrates the visualization paradigm of fitting and graphing. For the singer heights, it has led us to a more powerful description of the data. The sample means fitted to the data summarize the locations of the distributions. The graphs of the residuals make it clear that the distributions of the residuals are homogeneous and well approximated by the normal distribution. The pooling of the residuals to characterize the variation for each voice part and the approximation of the pooled residual distribution by the normal increase the information in the description of the variation of the data. The final result is a convincing, quantified picture of the relationship between height and pitch interval.

2.6 *Log Transformation*

In 1960 Bela Julesz sprang an ingenious invention on the scientific world of visual perception — the random dot stereogram [56, 57]. An example, designed by Julesz, is shown in Figure 2.18. Each of the two images has the appearance of a collection of random dots. But when the images are looked at in stereo — which means that one image falls on the left retina and the other image falls on the right — a 3-D object is seen. In this case, the object is a diamond-shaped region floating above or below the plane of the page.



2.18 A diamond-shaped region is shown in 3-D by a random dot stereogram.

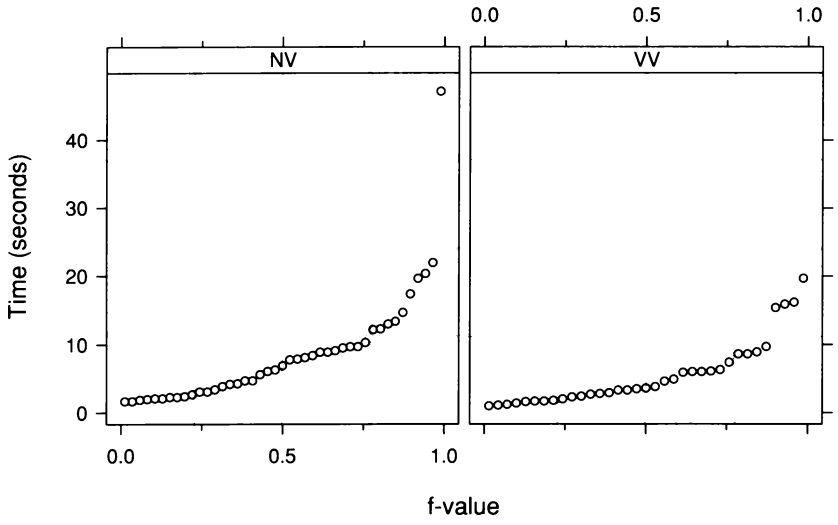
One way to fuse the two images in Figure 2.18 — that is, to see the 3-D effect — is to use a stereo viewer, which channels the left image to the left retina and the right image to the right retina. This makes the diamond float above the page. Another method requires no equipment — fixing on a point between the two images and defocusing the eyes by looking through the images, bringing them together in the middle. It is helpful to vary the distance of the image from the eyes to find an optimal position for fusion. When fusion is accomplished, the visual system actually sees three images, with the center one in 3-D, and, again, the diamond floats above the page. It is hard to make this method work the first time, but after a first success, it is typically easy to do again. A third method is crossing the eyes; since this sends the left image to the right retina and the right image to the left retina, the diamond floats below the page. For those without a stereo viewer, mastering the technique of viewing with the unaided eye will have a benefit. A visualization method in Chapter 4 uses stereo to display trivariate data.

The idea of the random dot stereogram is simple and elegant. Embedded in the right view is a region, in this case a diamond, that is exactly repeated in the left view, but is shifted slightly to the right. Quite incredibly, this *binocular disparity* is enough to enable fusion. Julesz's invention demonstrated that the visual system can process local detail to construct an object in 3-D, and needs no other information about the form of the object.

Typically, a viewer concentrating on a random dot stereogram achieves fusion after a few seconds or more. The fusion is not achieved by conscious thought, for example, by the processes that allow us to reason about the world. But fusing the stereogram once makes it easier to do again, so something stored in the brain's memory can decrease fusion time. An experiment was run to study the effect of prior knowledge of an object's form on fusion time [44]. The experimenters measured the time of first fusion for a particular random dot stereogram. There were two groups of subjects. The NV subjects received either no information or verbal information. The VV subjects received a combination of verbal and visual information, either suggestive drawings of the object or a model of it. Thus the VV subjects actually saw something that depicted the object, but the NV subjects did not. The goal in analyzing the fusion times is to determine if there is a shift in the distribution of the VV times toward lower values compared with the NV times. The experimenters used a classical method of probabilistic inference to analyze the data, and concluded that there is no shift. We will use visualization methods to re-examine this result.

Skewness

Figure 2.19 shows quantile plots of the fusion times. The data are *skewed toward large values*. Small values are tightly packed together, and the large values stretch out and cover a much wider range of the measurement scale. The extreme case is the largest NV time, 47.2 seconds, which is more than twice the value of the next largest observation. The skewing gradually increases as we go from small to large values; the result is a strongly convex pattern.

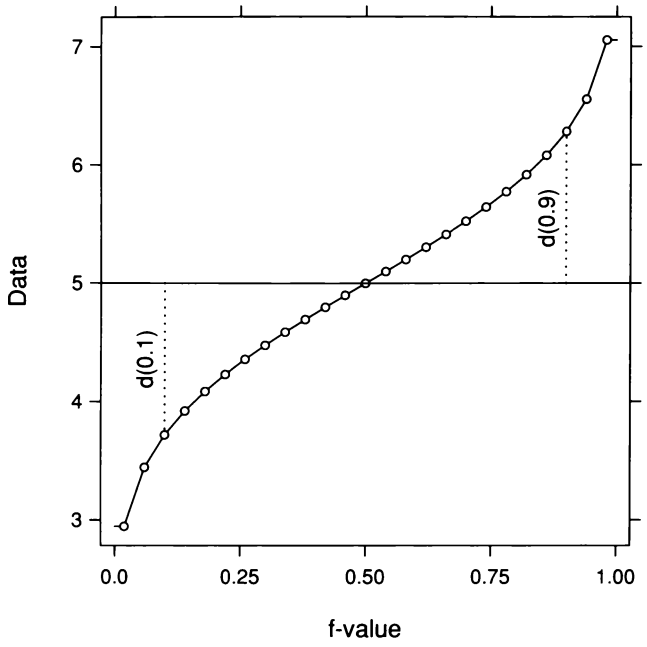


2.19 Quantile plots display the two distributions of the fusion-time data.

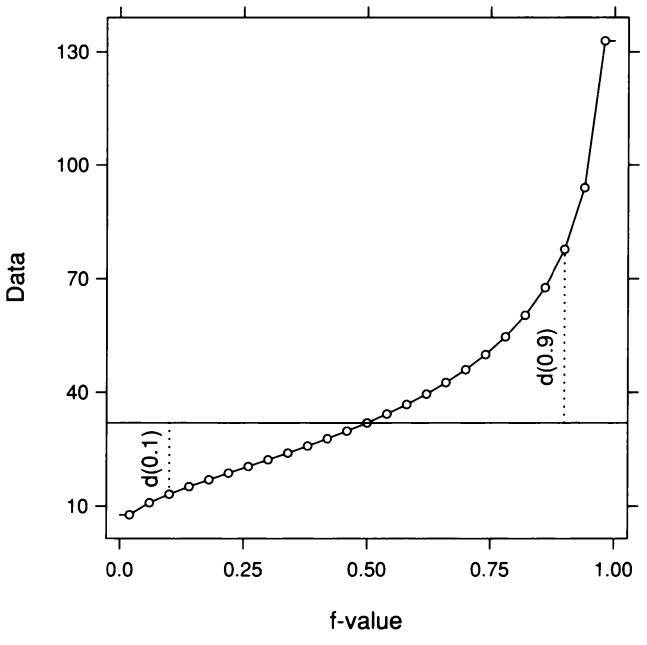
As before, let $q(f)$ be the f quantile of the data. Let $d(f)$ be the distance of $q(f)$ from the median,

$$d(f) = |q(0.5) - q(f)|.$$

A distribution is symmetric if $d(f)$ is symmetric about 0.5 as a function of f; that is, $d(f) = d(1 - f)$. The values shown on the quantile plot in Figure 2.20 are symmetric. A distribution is skewed toward large values if $d(f)$ is bigger than $d(1 - f)$ for f in the interval 0.5 to 1, and the disparity increases as f goes from 0.5 to 1, that is, from the center of the distribution to the tails. The values shown on the quantile plot of Figure 2.21 are skewed toward large values.

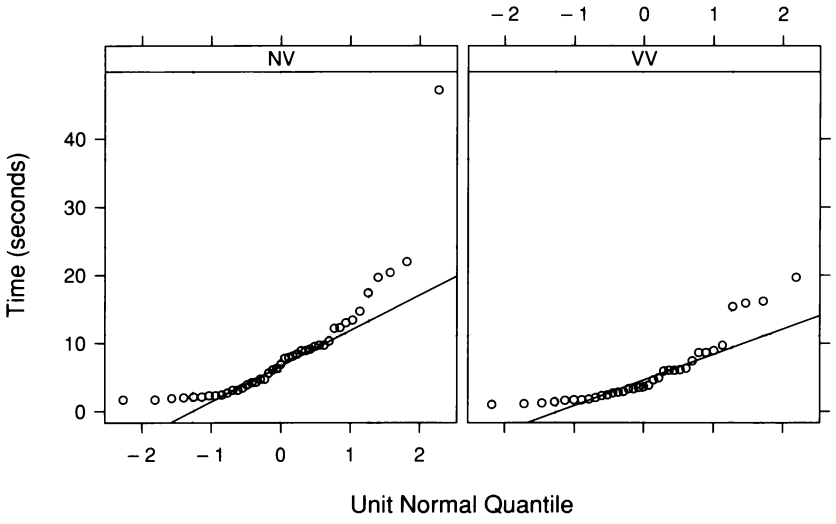


2.20 The graph shows the quantile function of data with a symmetric distribution.



2.21 The graph shows the quantile function of data that are skewed toward large values.

Figure 2.22 shows normal q-q plots of the fusion times. The skewness toward large values creates the same convex pattern as on the quantile plot. This deviation from normality occurs because the normal distribution is symmetric. This behavior contrasts with that of the distributions of the singer heights, which are nearly symmetric and are well approximated by the normal distribution.

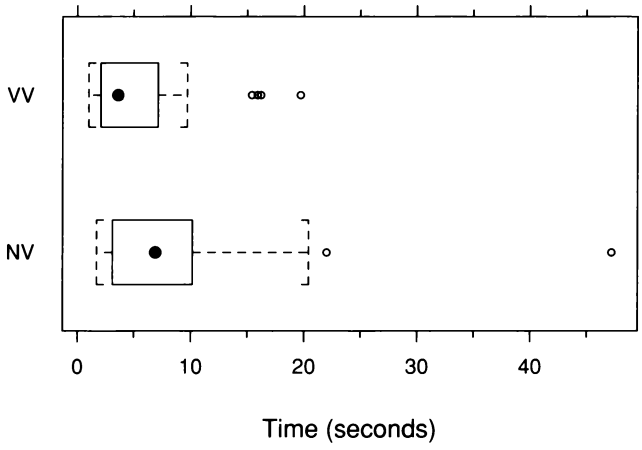


2.22 Normal q-q plots compare the distributions of the fusion-time data with the normal distribution.

Data that are skewed toward large values occur commonly. Any set of positive measurements is a candidate. Nature just works like that. In fact, if data consisting of positive numbers range over several powers of ten, it is almost a guarantee that they will be skewed. Skewness creates many problems. There are visualization problems. A large fraction of the data are squashed into small regions of graphs, and visual assessment of the data degrades. There are characterization problems. Skewed distributions tend to be more complicated than symmetric ones; for example, there is no unique notion of location and the median and mean measure different aspects of the distribution. There are problems in carrying out probabilistic methods. The distribution of skewed data is not well approximated by the normal, so the many probabilistic methods based on an assumption of a normal distribution cannot be applied. Fortunately, remedies coming in later sections can cure skewness.

Monotone Spread

In Figure 2.23, box plots of the fusion-time distributions show that the median of the NV times is greater than the median of the VV times, and the spread of the NV times is greater than the spread of the VV times. In other words, the spreads increase with the locations. The same phenomenon can also be seen on the normal q-q plots in Figure 2.22. The slope of the line on each panel is the interquartile range of the data divided by the interquartile range of the unit normal distribution. This measure of spread is just a rescaling of the interquartile range of the data. In Figure 2.22, the slope for the NV times is greater than the slope for the VV times.



2.23 The distributions of the fusion times are compared by box plots.

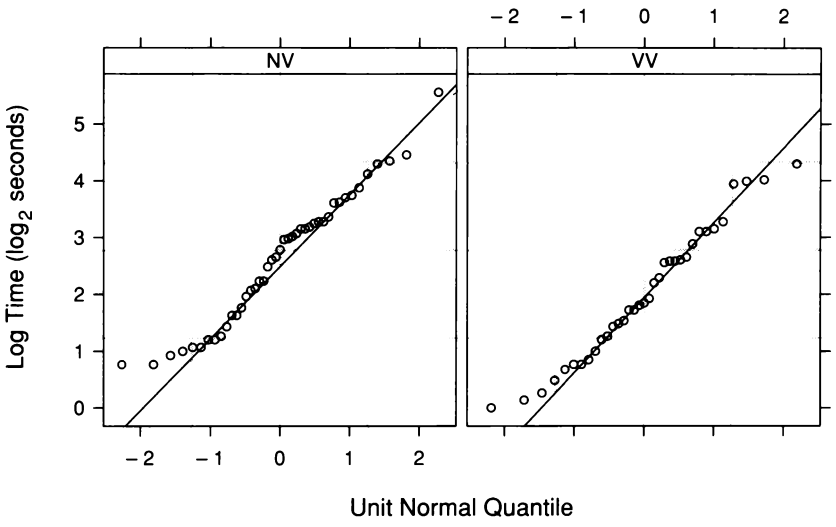
When the distributions of two or more groups of univariate data are skewed, it is common to have the spread increase monotonically with location. This behavior is *monotone spread.* Strictly speaking, monotone spread includes the case where the spread decreases monotonically with location, but such a decrease is much less common for raw data. Monotone spread, as with skewness, adds to the difficulty of data analysis. For example, it means that we cannot fit just location estimates to produce homogeneous residuals; we must fit spread estimates as well. Furthermore, the distributions cannot be compared by a number of standard methods of probabilistic inference that are based on an assumption of equal spreads; the standard t-test is one example. Fortunately, remedies for skewness can cure monotone spread as well.

Transformation by Logs

For positive univariate measurements, it is often more natural to consider multiplicative effects rather than additive ones. The fusion times are one example. Consider times of 4 seconds and 6 seconds for two individuals. The interesting information is that the second individual took 1.5 times as long. The two individuals differ by a lot more in their fusion performance than two individuals with times of 20 seconds and 22 seconds. Even though the absolute difference, 2 seconds, is the same in both cases, the poorer performer in the first case took 50% longer, whereas the poorer performer in the second set took only 10% longer. Taking logs amounts to changing the units of the data — from seconds to log seconds for the fusion times — in such a way that equal differences now mean equal multiplicative factors. This simplifies the interpretation of the measurement scale because, to put it simplistically, addition is easier than multiplication.

The logarithm is one of many transformations that we can apply to univariate measurements. The square root is another. Transformation is a critical tool for visualization or for any other mode of data analysis because it can substantially simplify the structure of a set of data. For example, transformation can remove skewness toward large values, and it can remove monotone increasing spread. And often, it is the logarithm that achieves this removal.

Figure 2.24 shows normal q-q plots of the logs of the fusion times. The data are now much closer to symmetry, although there is a small amount of remaining skewness; the points in the lower tail of each distribution lie somewhat above the line. In other words, the data quantiles in the lower tail are a bit too big, or too close to the median. The effect is small, but there is a plausible explanation. There is a minimum amount of time that a subject needs to consciously realize that the visual system has fused the stereogram, and then to signal that fusion has occurred. Effectively, this makes the origin somewhat bigger than 0 seconds. We could try to bring the lower tail into line by subtracting a small constant such as 0.5 seconds from all of the data, but this gets fussier than we need to be.



2.24 Normal q-q plots compare the distributions of the log fusion times with the normal distribution.

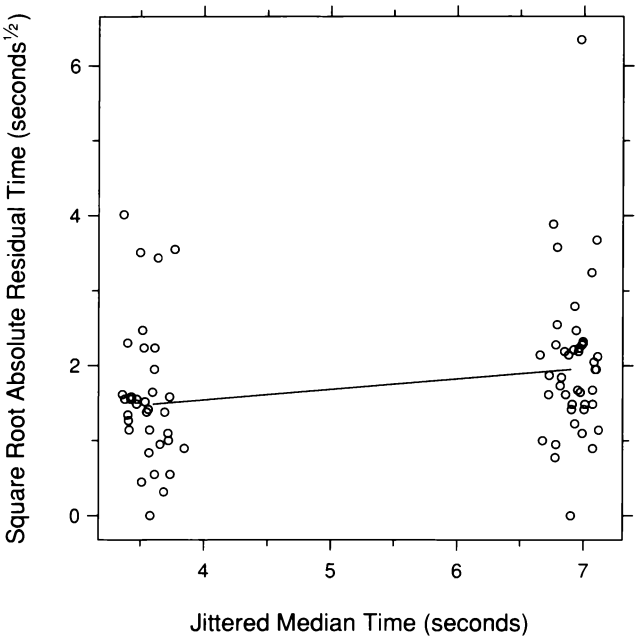
In graphing the log fusion times, log base 2 has been used. Thus, a change of one unit on the transformed scale means a change on the original scale by a factor of two. When data range through just a few powers of 10, $\log_2$ is easier to interpret than $\log_{10}$ because fractional powers of 10 are harder to fathom than integer powers of 2 [20].

Except for the outlier, the fusion times range from 1 second to 22 seconds. On the $\log_2$ scale, this range becomes 0 $\log_2$ seconds to 4.5 $\log_2$ seconds. On the $\log_{10}$ scale, the range becomes 0 $\log_{10}$ seconds to 1.3 $\log_{10}$ seconds, and we must expend effort fathoming fractional powers of 10 to comprehend the multiplicative effects.

S-L Plots

Monotone spread, when it occurs, can typically be spotted on box plots. For example, Figure 2.23 revealed the monotone spread of the fusion times. But the box plot is a general-purpose visualization tool for exploring many aspects of distributions, not just spread. The *spread-location plot*, or *s-l plot*, provides a more sensitive look at monotone spread because it is a specialized tool whose sole purpose is to detect changes in spread. The s-l plot translates looking at spread to looking at location. First, medians are fitted to the distributions of the data, x_i. Measures of location for the absolute values of the residuals, $|\widehat{\varepsilon}_i|$, are measures of spread for the x_i. For example, the *median absolute deviations*, or *mads*, of the distributions of the x_i are the medians of the distributions of the $|\widehat{\varepsilon}_i|$ [51, 65].

Figure 2.25 is an s-l plot of the fusion times. The circles graph the square roots of the $|\widehat{\varepsilon}_i|$ for the two distributions against the fitted values, which take on two values, the medians of the two distributions. To prevent undue overlap of the plotting symbols, the locations of the symbols are jittered by adding uniform random noise to the fitted values. Jittering will discussed further in Chapter 3.

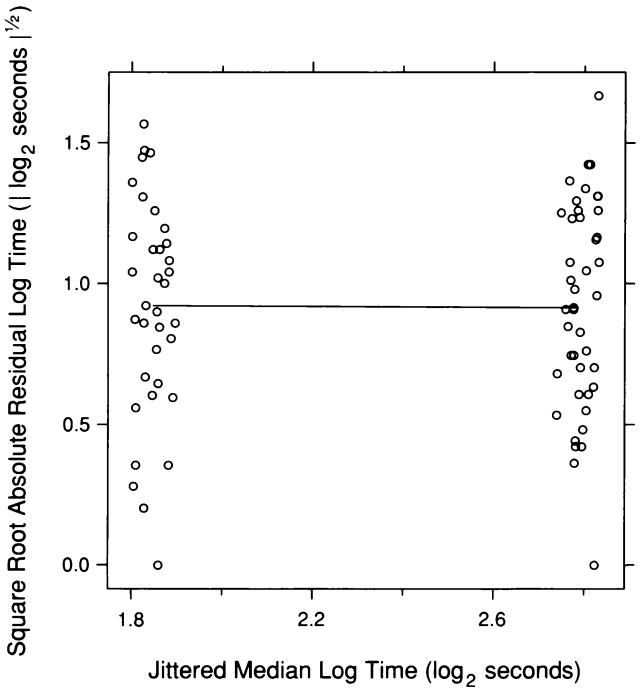


2.25 The s-l plot for the fusion times checks for nonuniform spread.

The square root transformation is used because absolute residuals are almost always severely skewed toward large values, and the square root often removes the asymmetry. Also, the square roots of the two mads are graphed against the two medians by connecting the plotting locations by a line segment.

The s-l plot for the fusion times adds to our understanding of the monotone spread of the distributions. It shows a convincing upward shift in the location of the $\sqrt{|\widehat{\varepsilon_i}|}$ for the NV times.

Figure 2.26 is an s-l plot for the log fusion times. The log transformation, in addition to removing most of the skewness, makes the spreads nearly equal. It is not unusual for a single transformation to do both. Nature is frequently kind enough to allow this.

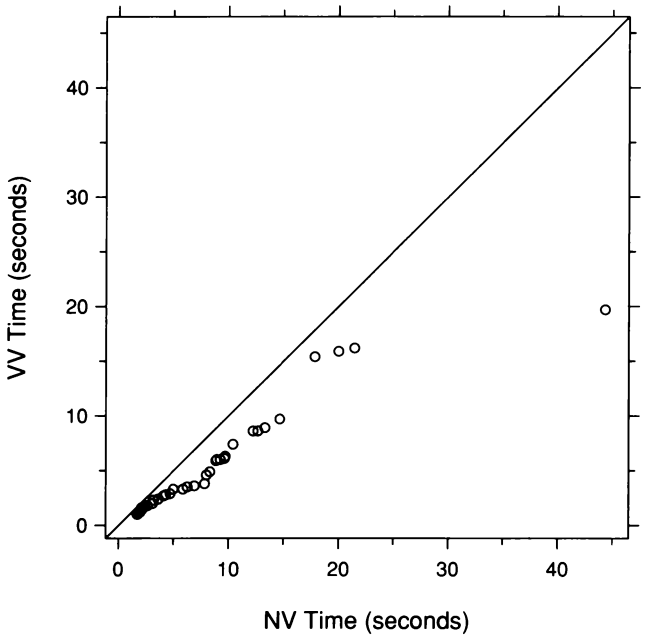


2.26 The s-l plot for the log fusion times checks for nonuniform spread.

Multiplicative and Additive Shifts

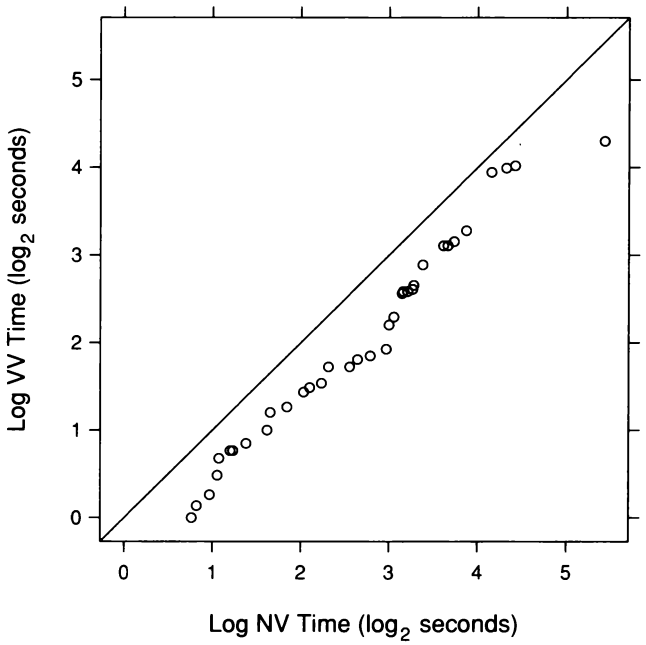
As we saw for the fusion times, logs can simplify structure by changing severe skewness to near symmetry and by changing monotone spread to nearly equal spread. It can also simplify structure by changing multiplicative shifts among distributions to additive shifts.

Figure 2.27 is a q-q plot of the fusion times on the original scale, before taking logs. There is a shift in the two distributions, but unlike the singer heights, it is not additive. The underlying pattern is a line through the origin with a slope of about 2/3. The shift is multiplicative, and quantiles with large f-values differ by more than those with small ones. This multiplicative shift is more complicated than an additive one because it results in a difference not just in the locations of the distributions but in the spreads as well. This produces the monotone spread.

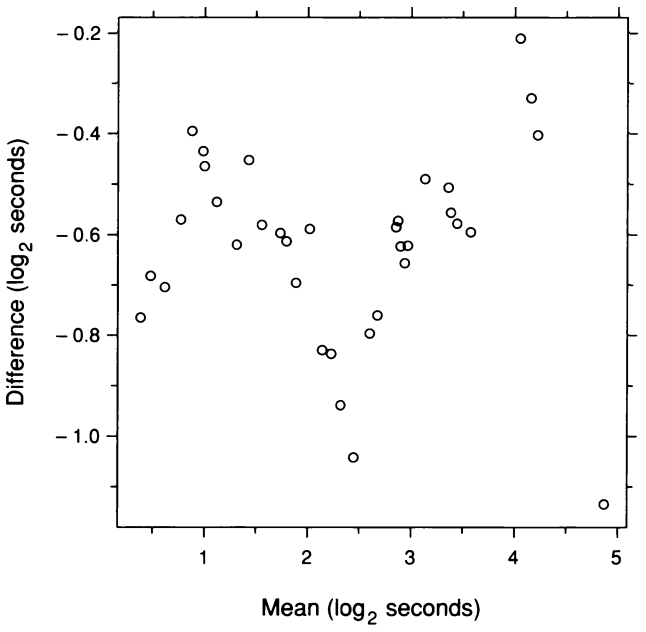


2.27 The q-q plot compares the distributions of the fusion times.

Figure 2.28 is a q-q plot of the log fusion times, and Figure 2.29 is a corresponding m-d plot. Now, on the log scale, the effect is additive. On the average, the log VV times are about 0.6 $\log_2$ seconds less than the log NV times. Back on the original scale, this is just the multiplicative effect with a multiplicative constant of $2^{-0.6} = 0.660 \approx 2/3$.



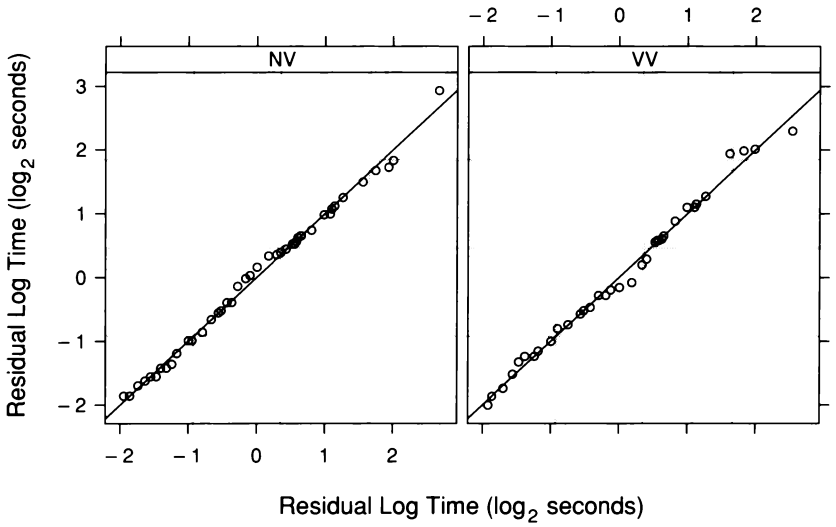
2.28 The q-q plot compares the distributions of the log fusion times.



2.29 The m-d plot provides more information about the shift of the log fusion times.

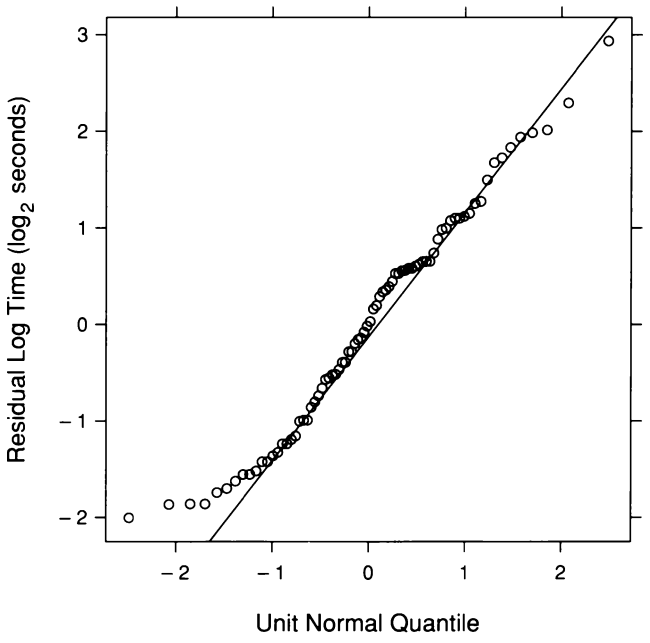
Fitting and Residuals

Since logs result in an additive shift of the fusion times, the fitting on the log scale only needs to account for a shift in location. Since the distributions are not far from the normal, means can be used to estimate locations. The mean for the log NV times is 2.6 $\log_2$ seconds, and the mean for the log VV times is 2.0 $\log_2$ seconds. Figure 2.30 shows q-q plots of the residuals; the quantiles of each residual distribution are graphed against the quantiles of the pooled residuals. The patterns lie close to the lines, which have intercept 0 and slope 1, so the two residual distributions are nearly the same. The residuals are homogeneous and can be pooled to characterize the variation in the data.

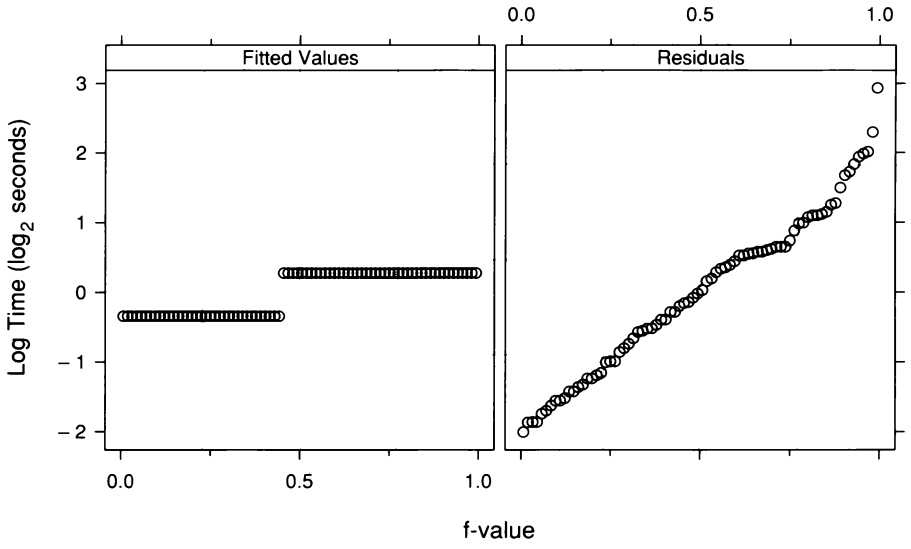


2.30 Each panel is a q-q plot that compares the distribution of the residuals for one group of log times with the distribution of the pooled residuals.

Figure 2.31, a normal q-q plot of the pooled residuals, shows the lifting of the lower tail observed earlier. The departure from normality begins at a value of −1 on the horizontal axis. This is the 0.16 quantile of the unit normal distribution. Thus the upper 84% of the distribution of the pooled residuals is well approximated by the normal. Figure 2.32, an r-f spread plot, shows that the spread of the residuals is considerably greater than the spread of the fitted values. Thus the effect of the increased VV information is small compared with other factors that affect fusion time.



2.31 A normal q-q plot compares the normal distribution with the distribution of the pooled residuals for the fit to log fusion time.



2.32 An r-f spread plot compares the spreads of the residuals and the fitted values minus their mean for the fit to log fusion time.

In the original analysis of the fusion-time data, the experimenters concluded that there was no support for the hypothesis that prior information about the object reduces fusion time. Yet our analysis here does seem to suggest an effect: the VV distribution is shifted toward lower values by a factor of 2/3. This apparent conflict between the original analysis and the analysis here is discussed in the final section of this chapter where it is argued that our visualization of the data has yielded valid insight that was missed by the experimenters.

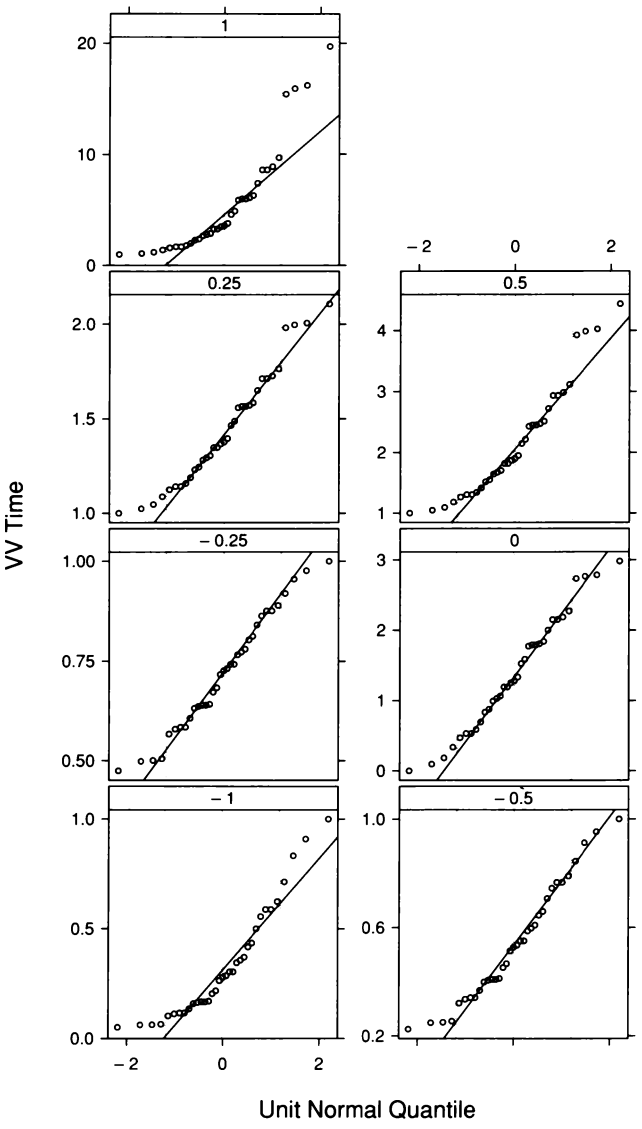
2.7 *Power Transformation*

We need more than just the logarithm in our arsenal of transformation methods. Logs cannot be used for data with zeros unless the data are adjusted. And a transformation other than the logarithm is often the one that leads to a simple data structure; in many cases the logarithm can fail to cure skewness and monotone spread, but another transformation does so.

Power transformations are a class of transformations that includes the logarithm. Let x be the variable under study. The power transformation with parameter τ is defined to be x^τ if $\tau \neq 0$ and $\log(x)$ if $\tau = 0$. For $\tau = 1$, the transformation is just x, so this leaves the data alone. For $\tau = 1/2$, the transformation is the square root, $\sqrt{x}$. For $\tau = -1$, it is the inverse of the data, $1/x$. It might seem artificial to define the power transformation for $\tau = 0$ to be the logarithm, but in fact it belongs there because x^τ for τ close to zero behaves much like the logarithm; for example, the derivative of $\log x$ is x^{-1}, and the derivative of $x^{0.001}$ is proportional to $x^{-.999}$. The parameter τ can be any number if the data are positive, but τ must be greater than 0 if the data have zeros. Of course, if the data have negative values, no power transformation is possible without some adjustment of the data.

Figure 2.33 shows normal q-q plots of the VV fusion times transformed by seven power transformations with values of τ equal to -1, $-1/2$, $-1/4$, 0, $1/4$, $1/2$, and 1. The panels are in graphical order: the value of τ increases as we go from left to right and from bottom to top through the panels. The figure illustrates a phenomenon that occurs for many data sets that are skewed toward large values. As τ decreases from 1, the skewness is reduced until the data become nearly

symmetric, and then as τ is further reduced, the data become more and more skewed again. When τ goes from 1 to $1/2$, the transformation pushes the upper tail closer to the center and pulls the lower tail further from the center. This continues as τ decreases, although when τ goes negative, the transformation reverses the upper and lower tails, so the continued force of the transformation pulls out the upper tail and pushes in the lower.



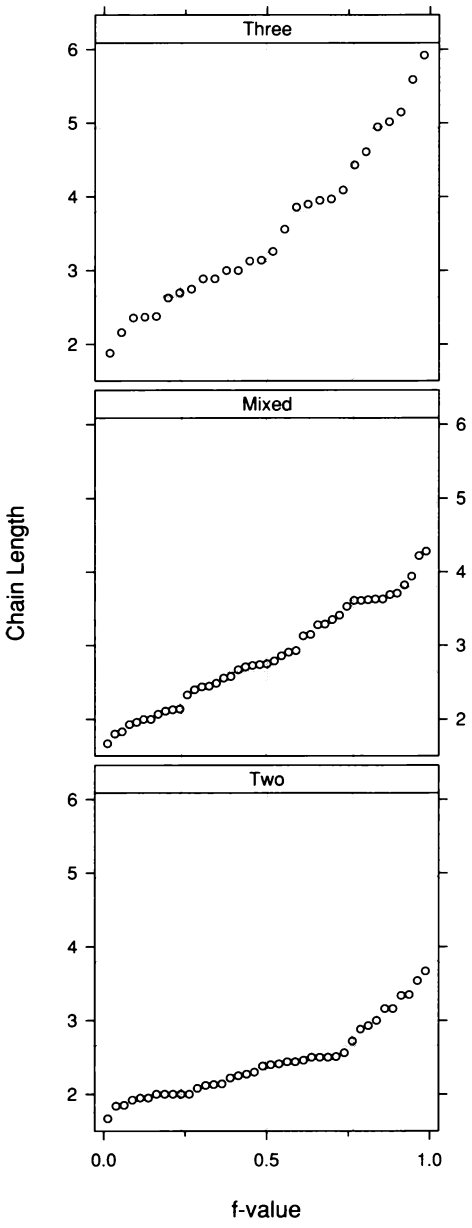
2.33 Seven power transformations of the VV times are displayed by normal q-q plots.

Figure 2.33 illustrates the method for discovering the power transformation that brings a distribution the closest to symmetry — trial and error. We simply choose a selection of values of τ, graph the transformed data for each transformation, and assess the symmetry. The seven values of τ in Figure 2.33 are a good starting set because they provide a representative collection of power transformations. For the VV times, the logarithm and the inverse fourth root do the best job. They also do the best job for the NV times. But, of course, a tie goes to the logarithm because of the multiplicative interpretation.

Food Webs

The food web for the animal species in an ecosystem is a description of who eats whom. A chain is a path through the web. It begins with a species that is eaten by no other, moves to a species that the first species eats, moves next to a species that the second species eats, and so forth until the chain ends at a species that preys on no other. If there are 7 species in the chain then there are 6 links between species, and the length of the chain is 6. The mean chain length of a web is the mean of the lengths of all chains in the web.

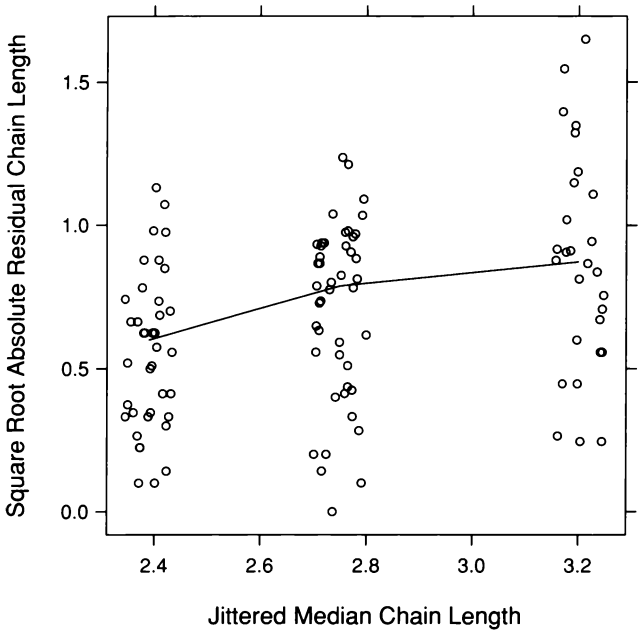
A two-dimensional ecosystem lies in a flat environment such as a lake bottom or a grassland; movement of species in a third dimension is limited. In a three-dimensional ecosystem, there is considerable movement in three dimensions. One example is a forest canopy; another is a water column in an ocean or lake. A mixed ecosystem is made up of a two-dimensional environment and a three-dimensional environment with enough links between the two to regard it as a single ecosystem. An interesting study reports the mean chain lengths for 113 webs [11]. Quantile plots display the data in Figure 2.34. Here, we will study how the distributions of mean chain lengths vary for the three classes — two-dimensional, mixed, and three-dimensional. In doing this, we regard the dimensionality of the mixed webs as lying between that of the two-dimensional webs and the three-dimensional webs.



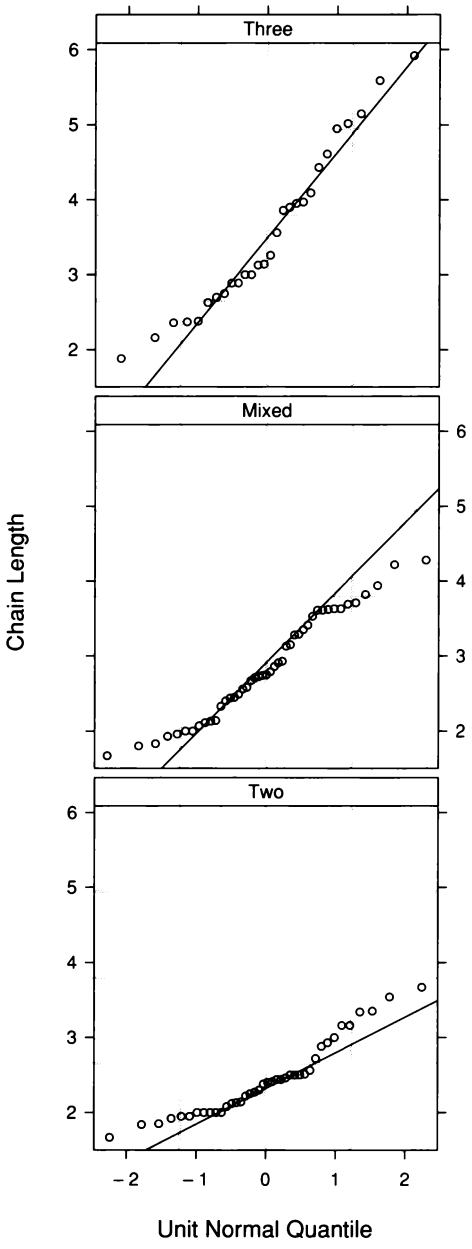
2.34 Quantile plots display the chain length measurements for three ecosystem dimensions.

Skewness and Monotone Spread

Figure 2.35 is an s-l plot of the three distributions of chain length. There is monotone spread. Normal q-q plots in Figure 2.36 reveal mild skewness toward large values. Also, the middle panel shows a peculiarity in the upper tail of the webs of mixed dimension. At a length of about 3.5, there is a knee in the distribution caused by five webs with very nearly identical values.

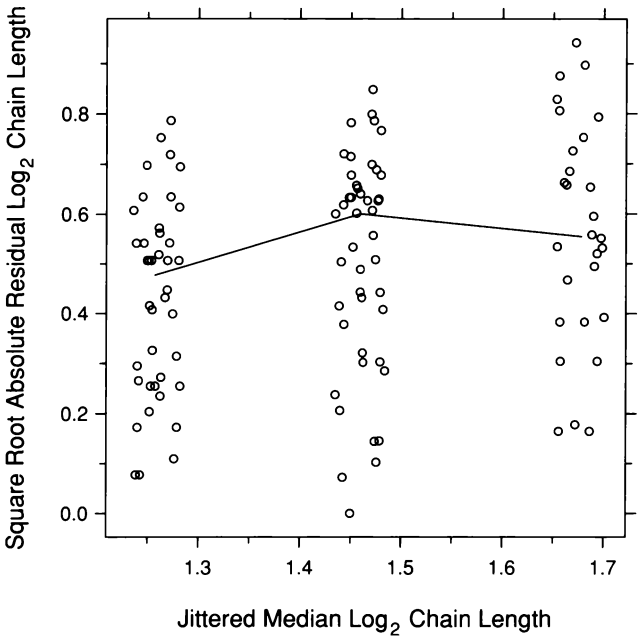


2.35 The s-l plot for the chain lengths checks for nonuniform spread.

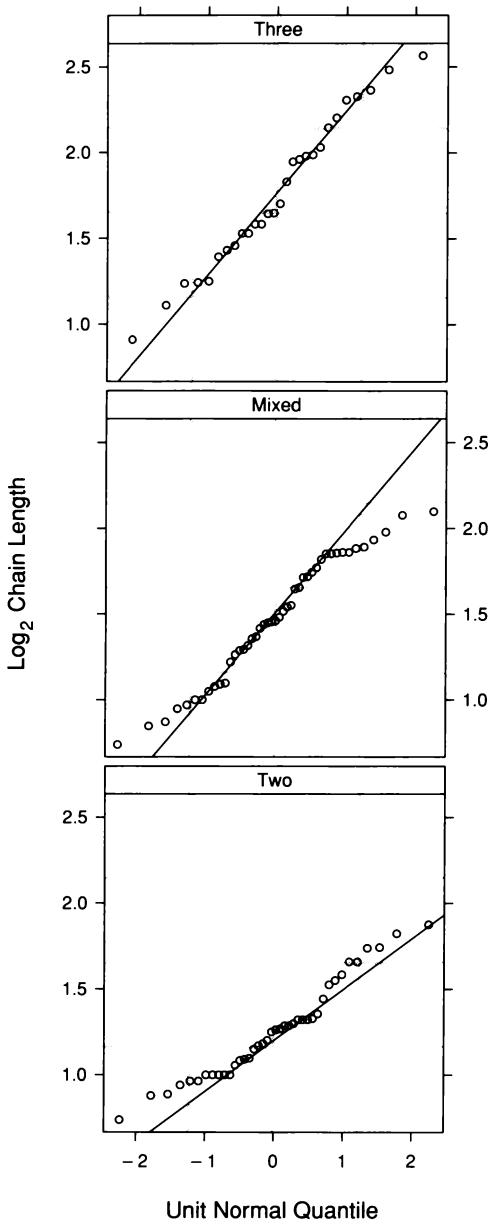


2.36 Normal q-q plots compare the chain length distributions with the normal distribution.

The skewness and monotone spread of the food web data are not cured by the logarithm, although the transformation does reduce their severity. Figure 2.37 is an s-l plot for the logs. The monotone spread remains, although the magnitude has been substantially reduced. Figure 2.38 shows normal q-q plots. Skewness remains, particularly for dimension two. We need a smaller value of τ.



2.37 The s-l plot for the log chain lengths checks for nonuniform spread.



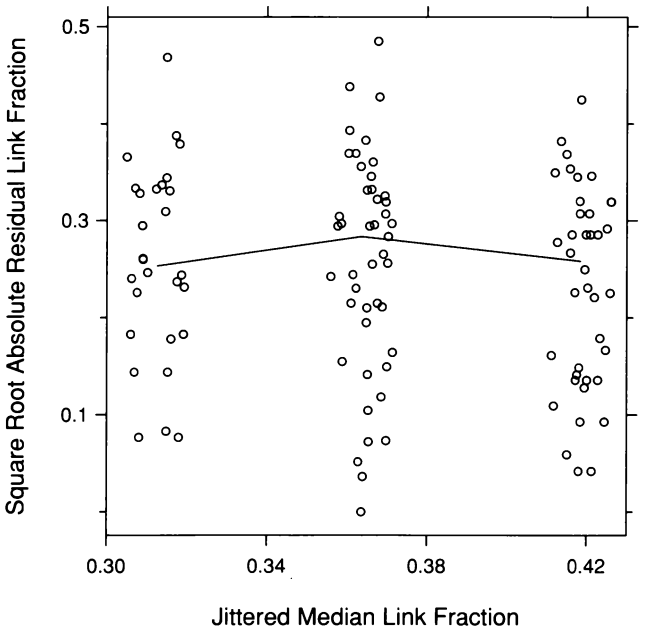
2.38 Normal q-q plots compare the log chain length distributions with the normal distribution.

The inverse transformation, which is $\tau = -1$, does the best job although the improvement over the logarithm is small. Figure 2.39 is the s-l plot for the inverse lengths; monotone spread no longer occurs. Figure 2.40 shows normal q-q plots of the inverses. The peculiar behavior for the mixed dimension now occurs in the lower tail because the inverse transformation has changed the order of the measurements. The other two panels, however, show a small amount of convexity in the lower tail of the distribution, so some of the peculiar behavior for the mixed dimension appears to be part of a general pattern.

The inverse transformation provides a natural measurement scale for the food web data. The measurement scale for chain length is links per chain. The measurement scale for inverse chain length is chains per link, or the *link fraction*. There is no reason to prefer links per chain to chains per link.

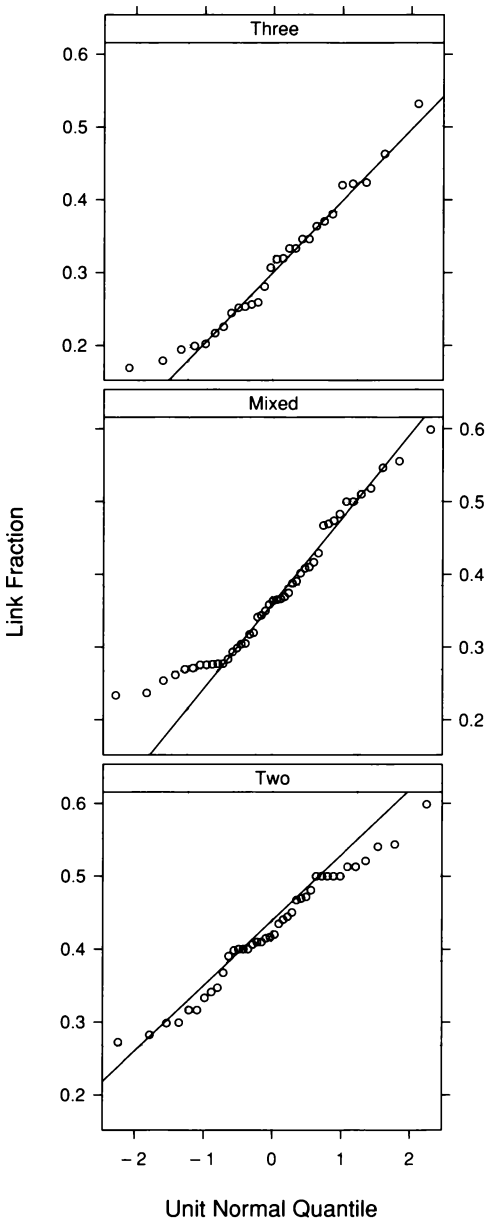
Fitting and Residuals

The normal q-q plots of the distributions in Figure 2.40 suggest that the shifts in the link-fraction distributions are additive on the inverse scale. Thus we will fit the data by estimating location, using means



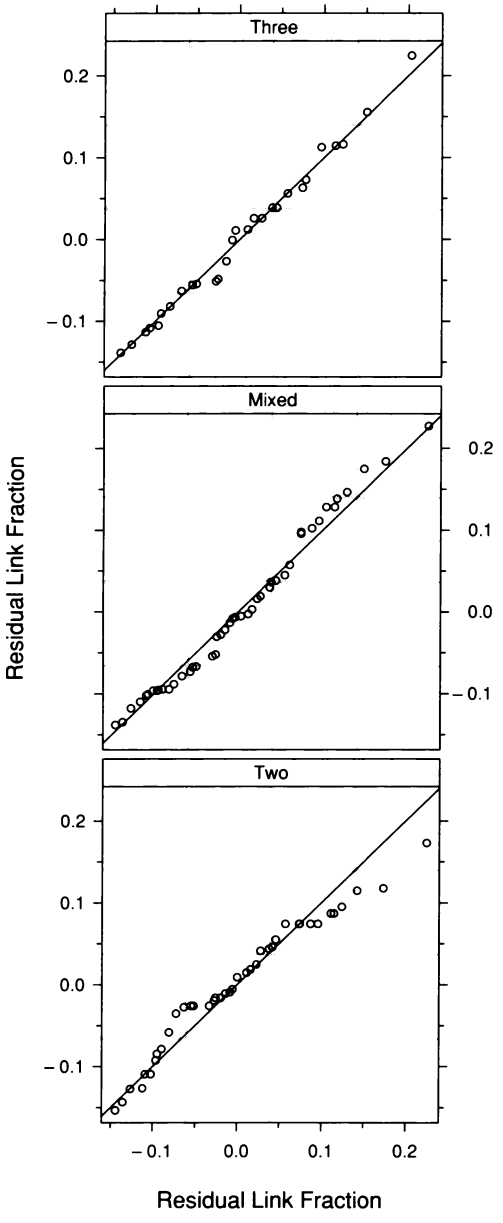
2.39 The s-l plot for the inverse chain lengths, or link fractions, checks for nonuniform spread.

because the distributions are not badly nonnormal. The mean link fractions are 0.43, 0.37, and 0.31 for two, mixed, and three. Note that these values are very close to 7/16, 6/16, and 5/16. Noting this form is likely just pure numerology, a taking of pleasure from detecting a simple pattern, but without theoretical significance.



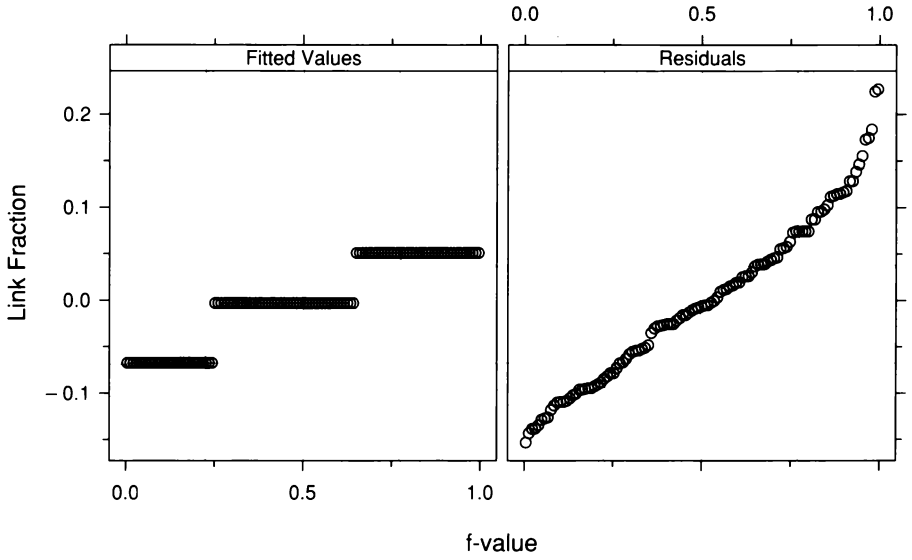
2.40 Normal q-q plots compare the normal distribution with the distributions of link fractions.

Figure 2.41 is a q-q plot of the residuals; the quantiles of each residual distribution are graphed against the quantiles of all residuals. The distributions are reasonably similar, so we can pool them.



2.41 Each panel is a q-q plot that compares the distribution of the residuals for one group of link fractions with the distribution of the pooled residuals.

Figure 2.42, an r-f spread plot, shows that the residual spread is considerably greater than the fitted-value spread. Thus, while the ecosystem dimension does affect the link fraction, it accounts for only a modest amount of the variation in the fractions.



2.42 An r-f spread plot compares the spreads of the residuals and the fitted values minus their mean for the fit to the link fractions.

The Force of Power Transformation

The amount of pushing and pulling on the tails of a distribution that is exerted by power transformations can be roughly measured by the ratio of the largest observation to the smallest. Data sets with large ratios are more sensitive to power transformation than data sets with small ratios. For the stereogram data, the ratio is 27.8 for the NV times and 19.7 for the VV times. For the food web data, it is 2.2 for two dimensions, 2.6 for mixed, and 3.1 for three. Power transformation clearly affects these two sets of data. If the ratio is too close to 1, power transformations with τ from -1 to 1 do not have much effect. For the eight singer distributions, the ratios range from 1.1 to 1.2, and power transformation has little effect on the shape of the distributions.

2.8 Complex Shifts and Robust Fitting

The visualization of the singer heights showed that the distributions differ by additive shifts, a simple structure that allows fitting by estimating location to produce homogeneous residuals. And the visualization showed that the residuals have a near-normal distribution. The discovery of power transformations for the fusion times and food web lengths resulted in measurement scales for which the distributions of these data sets also have additive shifts and near normality. Nature, however, is frequently not so obliging. Shifts between distributions can have an incurable complexity that makes them more difficult to characterize. And distributions can be severely nonnormal. Still, fitting and graphing can yield visualizations that provide a clear picture of the complexity and nonnormality.

Bin Packing Data

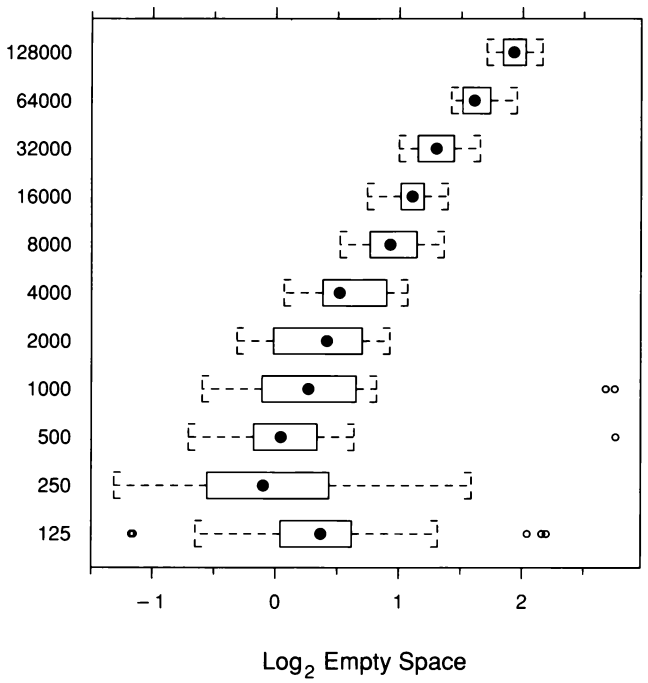
Bin packing is a computer problem that has challenged mathematicians working on the foundations of theoretical computer science. Suppose a large number of files of different sizes are to be written on floppies. No file can be split between two floppies, but we want to waste as little space as possible. Unfortunately, any algorithm that guarantees the minimum possible empty space takes an enormous amount of computation time unless the number of files is quite small. Fortunately, there are heuristic algorithms that run fast and do an extremely good job of packing, even though they do not guarantee the minimum of empty space. One is *first fit decreasing*. The files are packed from largest to smallest. For each file, the first floppy is tried; if it has sufficient empty space, the file is written, and if not, the second floppy is tried. If the second file has sufficient space, the file is written and if not, the third floppy is tried. The algorithm proceeds in this way until a floppy with space, possibly a completely empty one, is found.

To supplement the theory of bin packing with empirical results, mathematicians and computer scientists have run simulations, computer experiments in which bins are packed with randomly generated weights. For one data set from one experiment [8], the weights were randomly selected from the interval 0 to 0.8 and packed in bins of size one. The number of weights, n, for each simulation run took one

of 11 values: 125, 250, 500, and so forth by factors of 2 up to 128000. There were 25 runs for each of the 11 different numbers of weights, which makes $25 \times 11 = 275$ runs in all. For each run of the experiment, the performance of the algorithm was measured by the total amount of empty space in the bins that were used. We will study log empty space to enhance our understanding of multiplicative effects.

Shifts in Location, Spread, and Shape

Figure 2.43 displays the bin packing data by box plots. The shifts in the 11 distributions are complex. The locations tend to increase with n, the number of weights. For example, the medians increase with n except for the two smallest values of n. The box plots also show that the spreads of the distributions tend to decrease as n increases. Both the interquartile ranges and the ranges of the adjacent values tend to decrease with n.

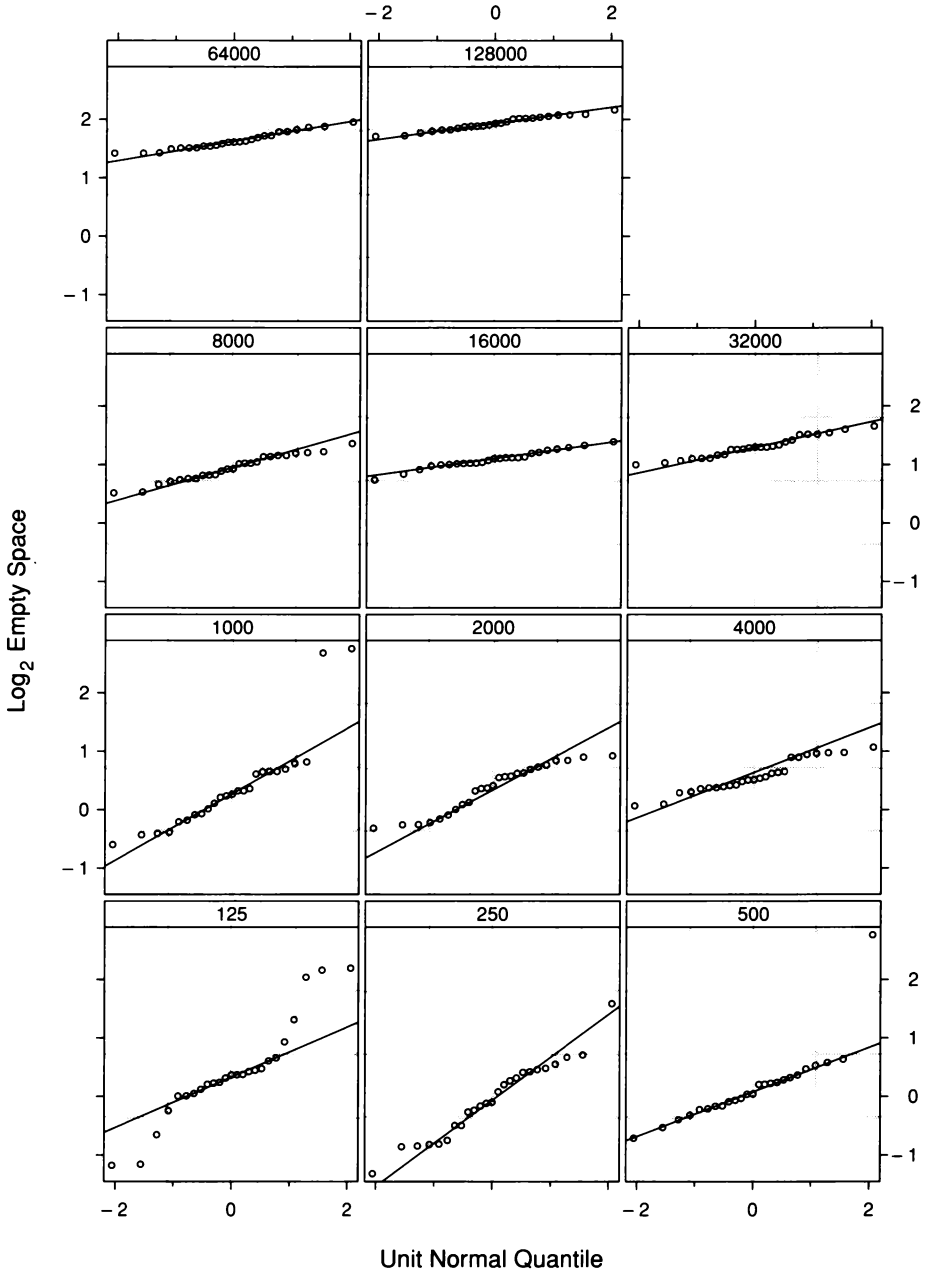


2.43 Box plots compare the distributions of the bin packing data.

Figure 2.44, normal q-q plots of the bin packing data, show that not all of the bin packing distributions are well approximated by the normal. For n equal to 2000 to 128000, the patterns are reasonably straight, but for lesser values of n, there are marked deviations from linear patterns. In other words, not only do the distributions differ in location and spread, they differ in shape as well. To study this shape in more detail, we will fit the data, compute residuals, and study the residual distributions. But since both location and spread change, we need to estimate both in fitting the data.

Figure 2.44 reveals enough information about shape to have a major impact on how we carry out the fitting of location and spread. For $n = 125$, the line follows the middle of the data; if the entire distribution were well approximated by the normal, the points in the tails would lie near this line. Instead, the points in the upper tail are above the line, and the points in the lower tail are below the line. In other words, with respect to a normal distribution as determined by the middle of the data, the values in the tails are too spread out. Distributions with this property occur frequently enough that there is a term to describe their shape — *leptokurtic*. The root "lepto" means slender. For data that are leptokurtic, the relative density of the data in the middle of the distribution compared with the density in the tails is less (thinner) than for a normal distribution. In cases where leptokurtosis affects only the extreme tails, the result can be just a few outliers. For n equal to 500 and 1000, three observations are considerably larger than the others. These outliers are likely just a remnant of a leptokurtosis that moves further and further into the tails of the distribution as n increases through 1000.

The opposite of leptokurtosis is *platykurtosis*. With respect to a normal distribution as determined by the middle of the data, the values in the tails of the data are a bit too close to the middle, that is, not sufficiently spread out. The root "platy" means broad. For data that are platykurtic, the relative density of the data in the middle of the distribution compared with the density in the tails is greater (broader) than for a normal distribution. Shortly, we will also see platykurtosis in the bin packing data.



2.44 Normal q-q plots compare the distributions of the bin packing data with the normal distribution.

Robust Estimation

For the singer heights, the log fusion times, and the link fractions, the distributions of the different groups of measurements are reasonably well approximated by the normal. For this reason, in fitting the data, we used sample means to estimate the locations of the distributions. Had it been necessary to account for changing spread, we would have used sample standard deviations to fit spread. Statistical theory tells us that sample means and standard deviations are the right thing to use when the distributions are close to the normal. And when a distribution is mildly platykurtic, sample means and standard deviations are typically satisfactory.

But in the presence of leptokurtosis, means and standard deviations can perform disastrously. Even just a few outliers can yield a silly answer. Suppose we have 20 observations ranging from 10 to 20 with a mean of 16.3. Suppose we add one new observation, an outlier equal to 99. The mean of the new data now becomes

$$(20 \times 16.3 + 99)/21 = 20.2 \, .$$

This is not a sensible description of the location of the data because it exceeds all of the observations except one. The sample standard deviation is also sensitive to a small number of outliers.

The median is typically *robust*: resistant to a few outliers or to more extreme leptokurtosis [51, 65]. Suppose that for the 20 observations above, the two middle observations — that is, the 10th and 11th largest observations — are 14 and 15. Then the median is 14.5. When the outlier is added, the median changes to 15, a sensible answer. Similarly, the mad, introduced earlier for use on s-l plots, is a robust estimate of spread [51, 65]. Thus we will use medians to fit the locations of the bin packing distributions, and mads to fit the spreads.

Let b_{in} be the ith log empty space measurement for the bin packing run with n weights. Let ℓ_n be the medians, and let s_n be the mads. The fitted values are

$$\widehat{b}_{in} = \ell_n \ ,$$

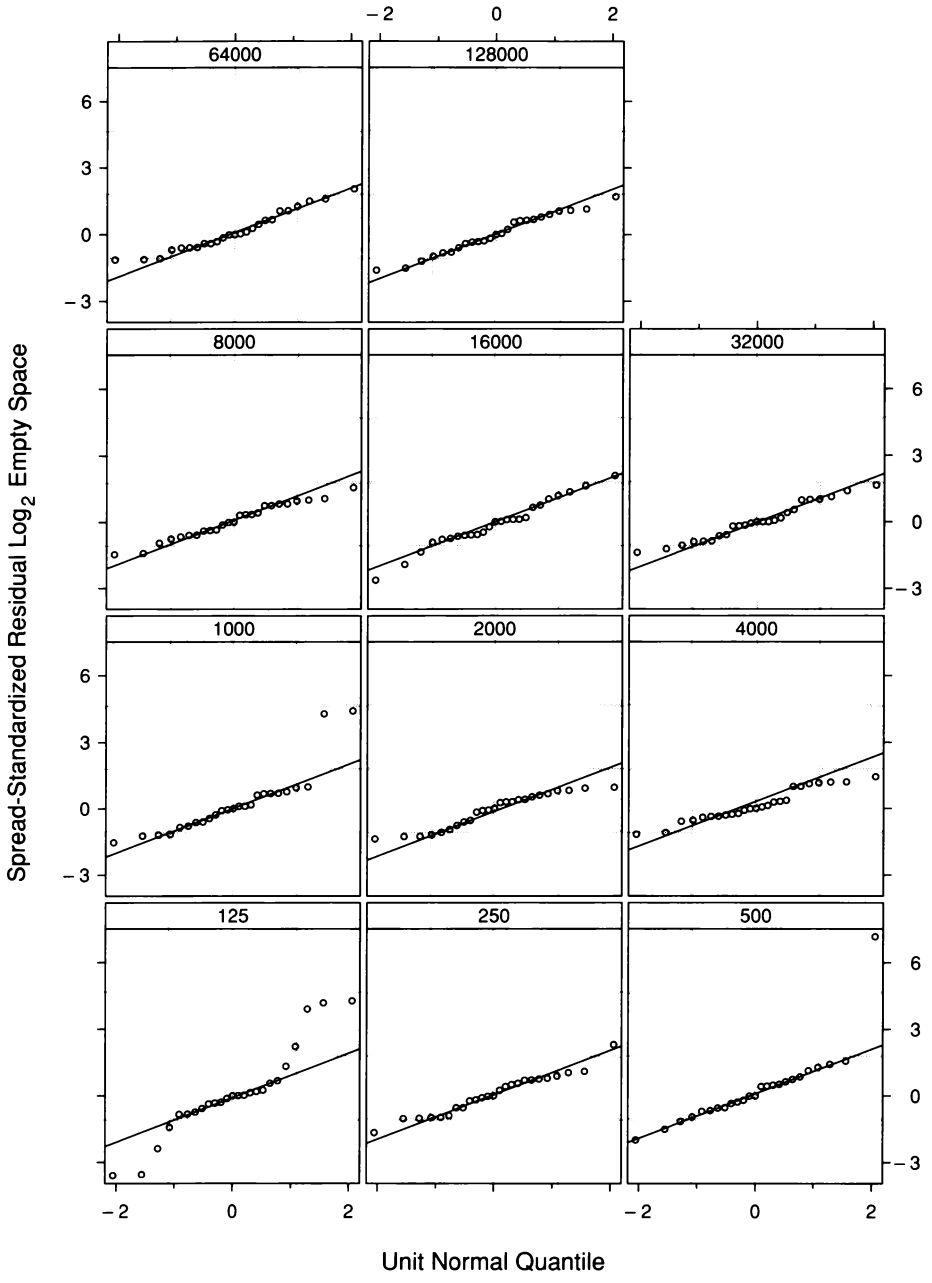
the residuals are

$$\widehat{\varepsilon}_{in} = b_{in} - \widehat{b}_{in} \ ,$$

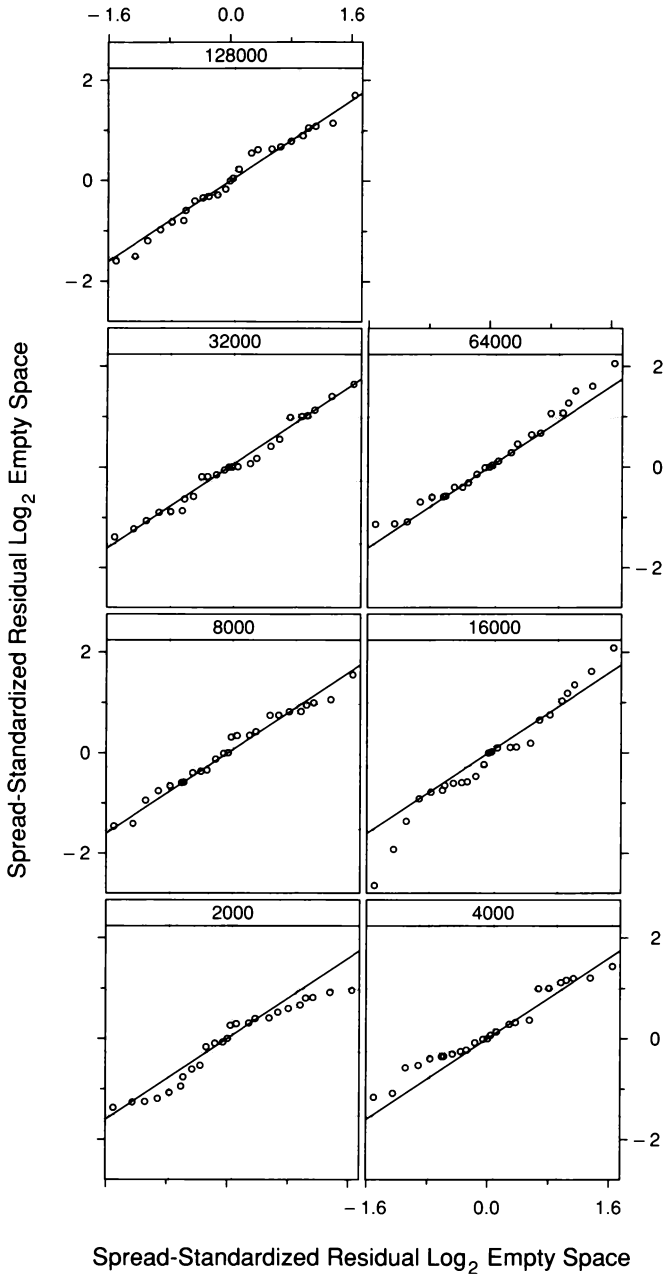
and the *spread-standardized residuals* are

$$\frac{\widehat{\varepsilon}_i}{s_n} \ .$$

In Figure 2.45, the distributions of the 11 sets of spread-standardized residuals are graphed by normal q-q plots. As before, the leptokurtosis is clearly revealed, but now we also see that once n is above 1000, the distributions turn mildly platykurtic. But a few outliers for distributions with $n \leq 1000$ squash the spread-standardized residuals for $n > 1000$ into a small region of the vertical scale, which interferes with our assessment of the platykurtosis. To enhance the assessment, we will eliminate the four distributions with $n \leq 1000$, and analyze the remaining seven residual distributions. Figure 2.46 shows q-q plots of each residual distribution against the pooled distribution of the seven sets. The departures for the three smaller values of n appear somewhat bigger than for the remaining, so there are differences among the distributions, but the effect is slight.

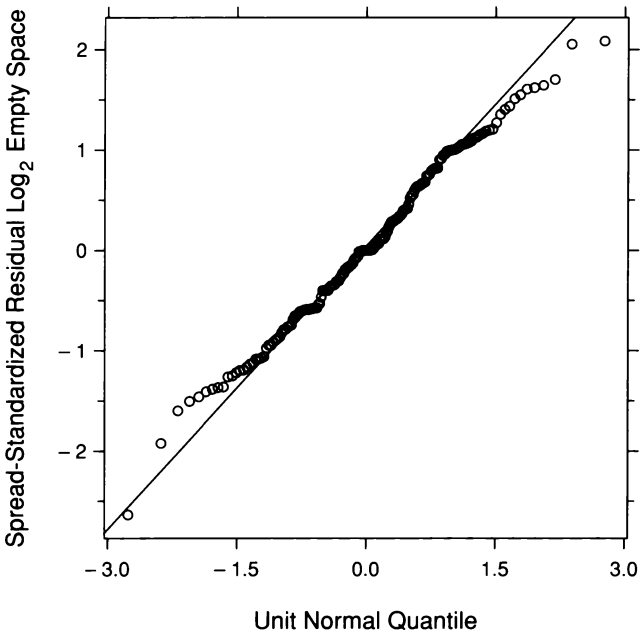


2.45 Normal q-q plots display spread-standardized residuals from the robust fit to the bin packing data.



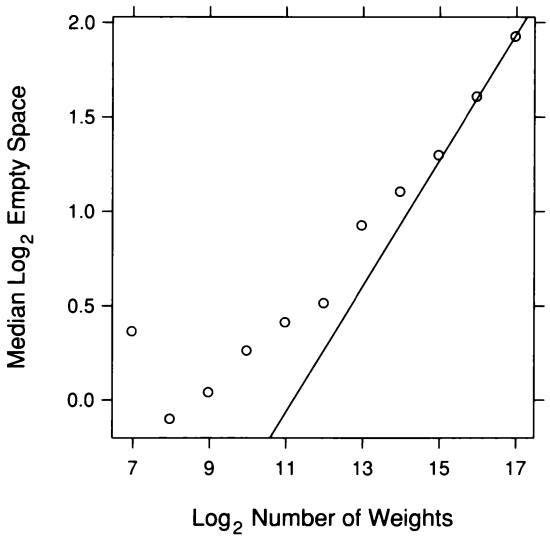
2.46 The q-q plots compare distributions of spread-standardized residuals for seven values of n with the pooled values of the seven distributions.

While Figure 2.46 has suggested some differences among the seven distributions with $n > 1000$, the magnitudes of the differences appear small, so we will pool the residuals anyway to further study the platykurtosis of the distributions. Figure 2.47, a normal q-q plot of the pooled values, shows that the shortening of the tails with respect to the normal begins near the quartiles of the data.



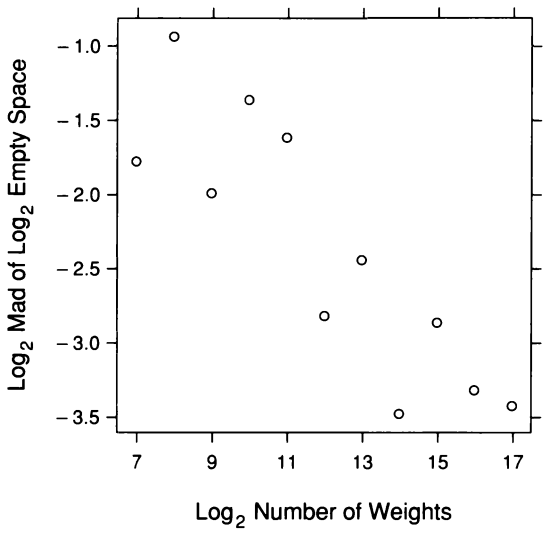
2.47 The normal q-q plot compares the normal distribution with the distribution of the pooled spread-standardized residuals for seven values of n.

Figure 2.48 graphs ℓ_n against $\log n$. Theoretical results suggest that when n gets large, ℓ_n is linear in $\log n$ with a slope of $1/3$ [8]. The line in Figure 2.48 has as slope of $1/3$ and passes through the rightmost point. The points for the largest values of $\log n$ do indeed appear to be increasing with this slope, but for the smaller values of $\log n$, the pattern of the points is convex and the slopes are less. In other words, before the asymptotics take over, the rate of growth in log empty space is less than $1/3$, and the rate increases as n increases.



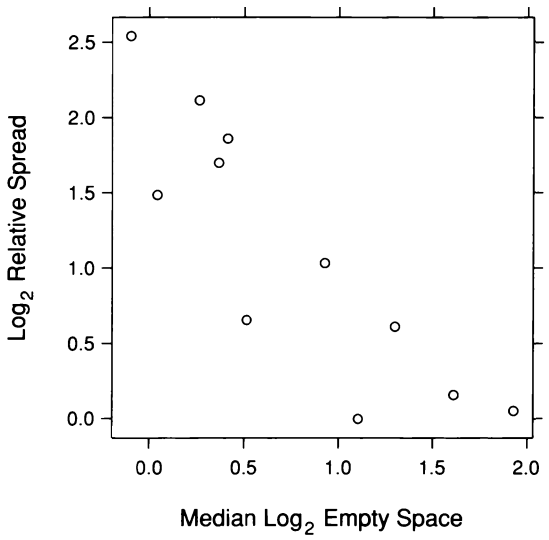
2.48 Median log empty space is graphed against log number of weights.

To study the behavior of s_n as a function of n, Figure 2.49 graphs log s_n against log n. The underlying pattern is linear, and a line with a slope of $-1/3$ would fit the pattern. This triggers an uncomfortable thought.



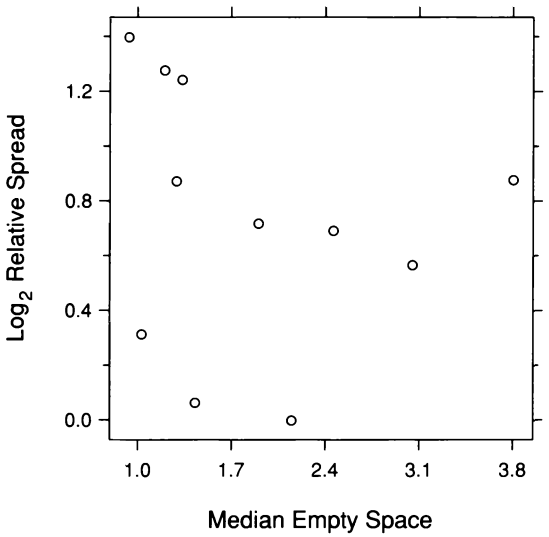
2.49 The logs of the mads for log empty space are graphed against log number of weights.

Since the log medians grow with a slope of $1/3$ as a function of $\log n$, and the log mads grow with a slope of $-1/3$ as a function of $\log n$, the log mads as a function of the log medians should have a slope of -1. This is checked in Figure 2.50; each s_n is divided by the minimum of the s_n, the log is taken, and the values are graphed against ℓ_n. This is simply an alternative form of the s-l plot; instead of graphing square root absolute residuals and square root mads against the medians, the log relative mads are graphed against the medians. The pattern is indeed linear with a slope of -1. In other words, there is monotone spread — a decrease in the spread with location. The discomfort is this. If the spreads of distributions do not depend on the locations, then taking logs can create monotone spread with exactly the pattern observed in Figure 2.50. We took the logs of the measurements of empty space at the outset to enhance the interpretation of multiplicative effects, but it is now likely that the transformation has induced the monotone spread.



2.50 Log relative mads for log empty space are graphed against median log empty space.

This is checked in Figure 2.51; the alternate s-l visualization method employed in Figure 2.50 is also used, but for empty space without transformation. Quite clearly, there is no monotone spread.



2.51 Log relative mads for empty space are graphed against median empty space.

Backing Up

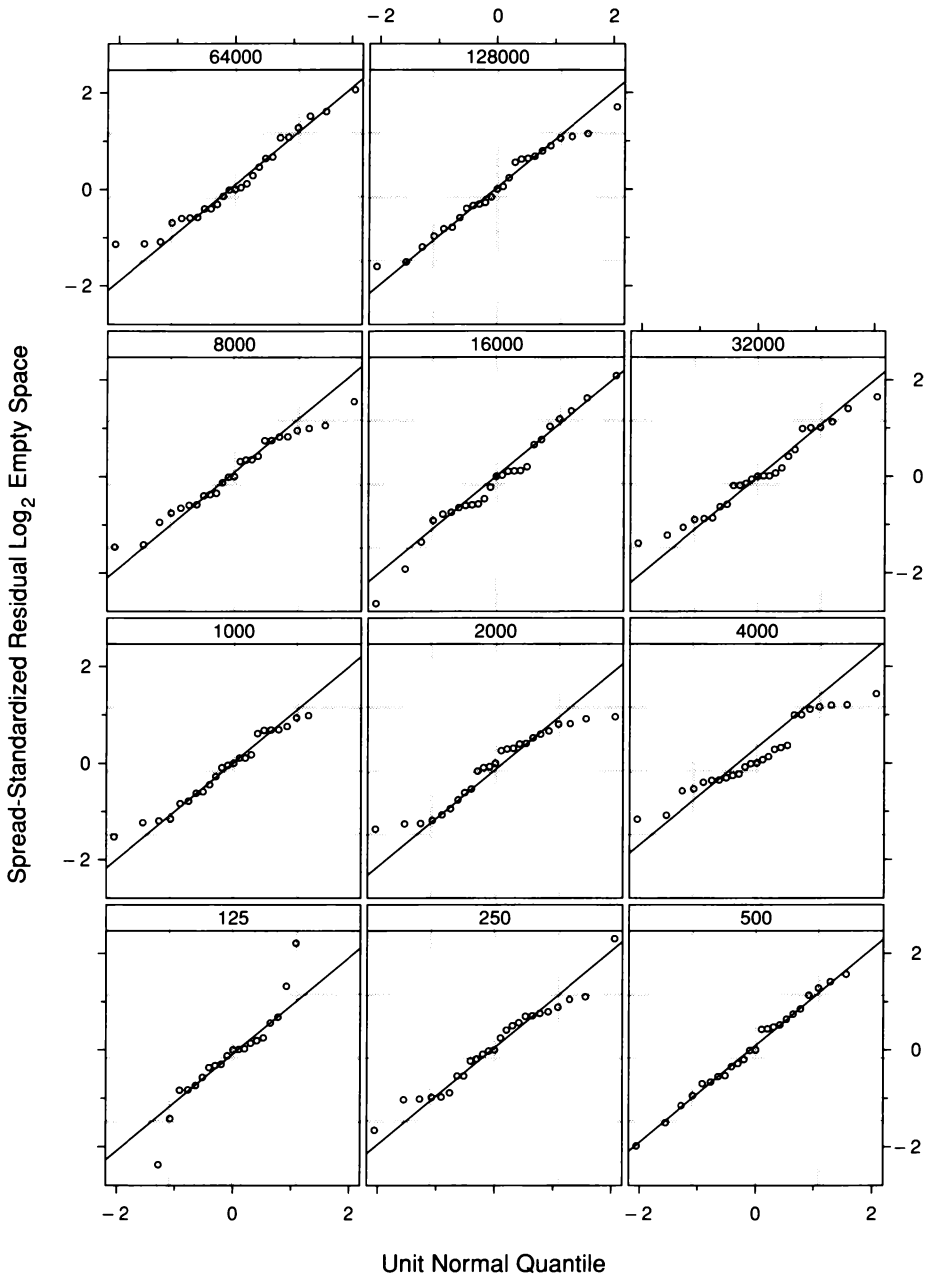
By taking logs at the outset to enhance the interpretation of multiplicative effects, we have induced some of the variation in the spreads of the distributions of empty space. With the benefit of visualization hindsight, it appears that we would be better off fitting empty space rather than log empty space. Then, we can always take logs of fits to study multiplicative properties. Thus, we should go back to the beginning and analyze the data again without transformation. We will not do so in this account since the data have been treated enough. Such backing up is common. It speaks to the power of visualization to show us mistaken actions that we have no way of judging until we visualize the results.

2.9 *Direct Manipulation: Outlier Deletion*

Outliers not only force us to use robust fitting, they ruin the resolution of graphical displays. A single large outlier that expands the range of the data by a factor of, say 2, squashes all of the remaining data into 50% of the scale on a graph. The normal q-q plot of the spread-standardized residuals in Figure 2.45 was the victim of such squashing. The range of all spread-standardized residuals is -3.59 to 7.17. But 97% of them lie between ± 2.75, so 3% of the spread-standardized residuals take up about 50% of the vertical scale. The result is a reduction in our ability to assess the nonnormality of the distributions aside from the outlier behavior. For example, the platykurtosis for n greater than 1000 is barely noticeable on the display. The remedy was to continue the visualization for the distributions with $n > 1000$. This removed the outliers. Another sensible procedure would have been to make Figure 2.45 again, but deleting the points whose vertical scale values lie beyond ± 2.75. The result of this outlier deletion is shown in Figure 2.52.

Direct manipulation provides an effective way to delete outliers. Since we identify outliers through the visualization process, it is natural to visually designate points for deletion by touching them with an input device. By their nature, outliers are typically few in number, so the process can be carried out quickly. By contrast, in computing environments that allow only static graphics, outlier deletion must be carried out by a less natural process. Scales must be read to determine cutoffs such as the ± 2.75 used for Figure 2.52, then commands must be issued to determine those observations that do not survive the cutoff, then more commands must be issued to delete them from the data given to the graphical routine, and finally the graph must be redrawn.

Outlier deletion by direct manipulation is so attractive that in 1969 Edward Fowlkes made it one of the first tools in the arsenal of direct manipulation methods discussed in Chapter 1. And outlier deletion was among the first operations of brushing, a direct manipulation tool that will be described in Chapter 3.



2.52 The outliers in the spread-standardized residuals have been removed. Direct manipulation provides a convenient environment for such point deletion.

2.10 Visualization and Probabilistic Inference

Visualization can serve as a stand-alone framework for making inferences from data. Or, it can be used to support probabilistic inferences such as hypothesis tests and confidence intervals. The fusion times provide a good example of this interplay between visualization and probabilistic inference.

The Julesz Hypothesis

Fusing a complicated random dot stereogram requires movements of the eyes to the correct viewing angle to align the two images. These movements happen very quickly in viewing a real scene; they bring focus to the scene as we shift our attention among objects at different distances. A more complex series of movements is needed to fuse a complicated random dot stereogram for the first time. But viewers can typically drastically reduce the time needed to fuse it by repeated viewing. Practice makes perfect because the eyes learn how to carry out the movements.

From informal observation, Bela Julesz noted that prior information about the visual form of an object shown in a random dot stereogram seemed to reduce the fusion time of a first look [57]. It certainly makes sense that if repeated viewing can reduce the fusion time, then seeing a 3-D model of the object would also give the viewer some information useful to the eye movements. This Julesz observation led to the experiment on fusion time, a rough pilot experiment to see if strong effects exist.

Reproducibility

Our visualization of the fusion times showed an effect; the distribution shifts toward longer times for reduced prior information. But the r-f plot showed the effect is small compared with the overall variation in fusion time. This makes it less convincing. Perhaps the better performance of increased information is spurious. Another experiment run in the same way might yield data in which decreased prior information produced shorter times. Once we begin thinking like this, it is time for probabilistic inference, not that we expect a definitive answer, but just simply to extract further information to help us judge the reproducibility of the results.

Statistical Modeling

But we must make the big intellectual leap required in probabilistic inference. We imagine that the subjects in the experiment are a random sample of individuals who can use a stereogram to see in depth. (A small fraction of individuals get no depth perception from stereo viewing.) Thus the log times for each group, NV and VV, are a random sample of the population of times for all of our specified individuals. Let μ_v be the population mean of the log VV times and let μ_n be the population mean of the log NV times.

Validating Assumptions

The normal q-q plots in our analysis showed that the two distributions of log times are reasonably well approximated by normal distributions. The lower tails are slightly elevated, but the departures are small. Thus it is reasonable to suppose that our two populations of log fusion times are well approximated by the normal distribution. The q-q plots and the fitting for viewing showed that it is reasonable to suppose the two population distributions have an additive shift. In particular, this means the standard deviations of the two populations are the same. With this checking of assumptions through visualization, we have earned the right to use the standard t-method to form a confidence interval for the difference in the two population means. Also, the visualization has revealed that it would be improper to use this method on the original data, which are severely nonnormal and do not differ by an additive shift.

Confidence Intervals

The sample mean of the 35 log VV times is $\bar{x}_v = 2.00 \log_2$ seconds. The sample mean of the 43 log NV times is $\bar{x}_n = 2.63 \log_2$ seconds. The estimate of the sample standard deviation from the residuals, $\widehat{\varepsilon}_i$, is

$$s = \sqrt{\frac{1}{76} \sum_{i=1}^{78} \widehat{\varepsilon}_i{}^2} = 1.18 \ \log_2 \text{ seconds} .$$

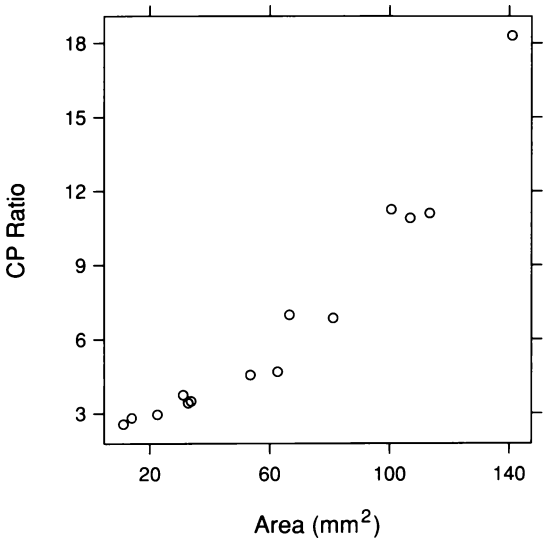
A 95% confidence interval for the difference in population means, $\mu_n - \mu_v$, is

$$(\bar{x}_n - \bar{x}_v) \pm 1.99 \, s \sqrt{\frac{1}{43} + \frac{1}{35}} ,$$

where 1.99 is the 0.975 quantile of a t-distribution with 76 degrees of freedom. The lower and upper values of the 95% interval are 0.09 and 1.15. Furthermore, the interval (0, 1.24) is a 97.7% interval. In other words, there is reasonable evidence that the results of the experiment are reproducible because a value of 0, which means no effect, does not enter a confidence interval until the confidence level is 97.7%.

Rote Data Analysis

The fusion-time experimenters based their conclusions on rote data analysis: probabilistic inference unguided by the insight that visualization brings. They put their data into an hypothesis test, reducing the information in the 78 measurements to one numerical value, the significance level of the test. The level was not small enough, so they concluded that the experiment did not support the hypothesis that prior information reduces fusion time. Without a full visualization, they did not discover taking logs, the additive shift, and the near normality of the distributions. Such rote analysis guarantees frequent failure. We cannot learn efficiently about nature by routinely taking the rich information in data and reducing it to a single number. Information will be lost. By contrast, visualization retains the information in the data.



3.1 A scatterplot displays bivariate data: measurements of retinal area and CP ratio for 14 cat fetuses.