3.9 For the lung cancer data set (see the codebook in Table 13.1 of Section 13.3 and Section A.7 of Appendix A for how to obtain the data) use a statistical package of your choice to
(a) compute a histogram of the variable Days, and
(b) for every other variable produce a frequency table of all possible values.

3.10 For the raw data in Problem 3.9
(a) produce a separate histogram of the variable Days for small and large tumor sizes (0 and 1 values of the variable Staget),
(b) compute a two-way frequency table of the variable Staget versus the variable Death, and
(c) comment on the results of (a) and (b).

3.11 For the depression data set, determine if any of the variables have observations that do not fall within the ranges given in Table 3.4, codebook for depression data.

3.12 For the statistical package you intend to use, describe how you would add data from three more time periods for the same subjects to the depression data set.

3.13 For the lung function data set, create a new variable called AGEDIFF = (age of child 1) – (age of child 2) for families with at least two children. Produce a frequency count of this variable. Are there any negative values? Comment.

3.14 Combine the results from the following two questions into a single variable:
a. Have you been sick during the last two weeks?

Yes, go to b. ___
No ___

b. How many days were you sick? ___

3.15 Consistency checks are sometimes performed to detect possible errors in the data. If a data set included information on sex, age, and use of contraceptive pill, describe a consistency check that could be used for this data set.

3.16 In the Parental HIV data set, the variable LIVWITH (who the adolescent was living with) was coded 1=both parents, 2=one parent, and 3=other. Transform the data so it is coded 1=one parent, 2=two parents, and 3=other using the features available in the statistical package you are using or Excel.

Chapter 4

Data screening and transformations

4.1 Transformations, assessing normality and independence

In Section 3.5 we discussed the use of transformations to create new variables. In this chapter we discuss transforming the data to obtain a distribution that is approximately normal. Section 4.2 shows how transformations change the shape of distributions. Section 4.3 discusses several methods for deciding when a transformation should be made and how to find a suitable transformation. An iterative scheme is proposed that helps to zero in on a good transformation. Statistical tests for normality are evaluated. Section 4.4 presents simple graphical methods for determining if the data are independent. In this chapter, we rely heavily on graphical methods: see Cook and Weisberg (1994) and Tufte (1997, 2001).

Each computer package offers the users information to help decide if their data are normally distributed. The packages provide convenient methods for transforming the data to achieve approximate normality. They also include output for checking the independence of the observations. Hence the assumption of independent, normally distributed data that is made in many statistical tests can be assessed, at least approximately. Note that most investigators will try to discard the most obvious outliers prior to assessing normality because such outliers can grossly distort the distribution.

4.2 Common transformations

In analysis of data it is often useful to transform certain variables before performing the analyses. Examples are found in the next section and in Chapter 6. In this section we present some common transformations. If you are familiar with this subject, you may wish to skip to the next section.

To develop a feel for transformations, let us examine a plot of transformed values versus the original values of the variable. To begin with, a plot of values of a variable X against itself produces a 45° diagonal line going through the origin, as shown in Figure 4.1.

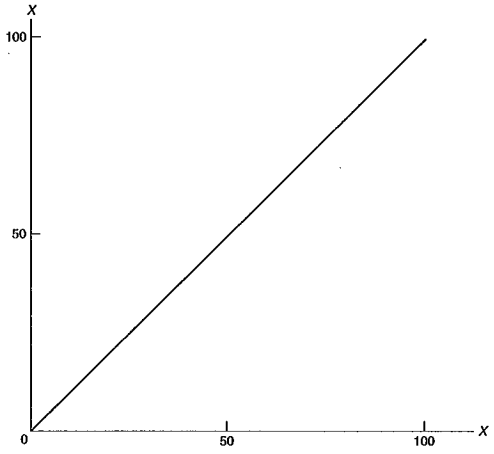


Figure 4.1: *Plot of Variable X versus Variable X*

One of the most commonly performed transformations is taking the **logarithm** (log) to base 10. Recall that the logarithm is the number that satisfies the relationship $X = 10^Y$. That is, the logarithm of X is the power Y to which 10 must be raised in order to produce X. As shown in Figure 4.2 in plot **a**, the logarithm of 10 is 1 since $10 = 10^1$. Similarly, the logarithm of 1 is 0 since $1 = 10^0$, and the logarithm of 100 is 2 since $100 = 10^2$. Other values of logarithms can be obtained from tables of common logarithms, from a hand calculator with a log function, or from statistical packages by using the transformation options. All statistical packages discussed in this book allow the user to make this transformation as well as others.

Note that an increase in X from 1 to 10 increases the logarithm from 0 to 1, that is, an increase of one unit. Similarly, an increase in X from 10 to 100 increases the logarithm also by one unit. For larger numbers it takes a great increase in X to produce a small increase in $\log X$. Thus the logarithmic transformation has the effect of stretching small values of X and condensing large values of X. Note also that the logarithm of any number less than 1 is negative, and the logarithm of a value of X that is less than or equal to 0 is not defined. In practice, if negative or zero values of X are possible, the investigator may first add an appropriate constant to each value of X, thus making them all positive prior to taking the logarithms. The choice of the additive constant can have an important effect on the statistical properties of the transformed variable, as will be seen in the next section. The value added must be larger than the magnitude of the minimum value of X.

Logarithms can be taken to any base. A familiar base is the number e = 2.7183 ... . Logarithms taken to base e are called **natural logarithms** and are

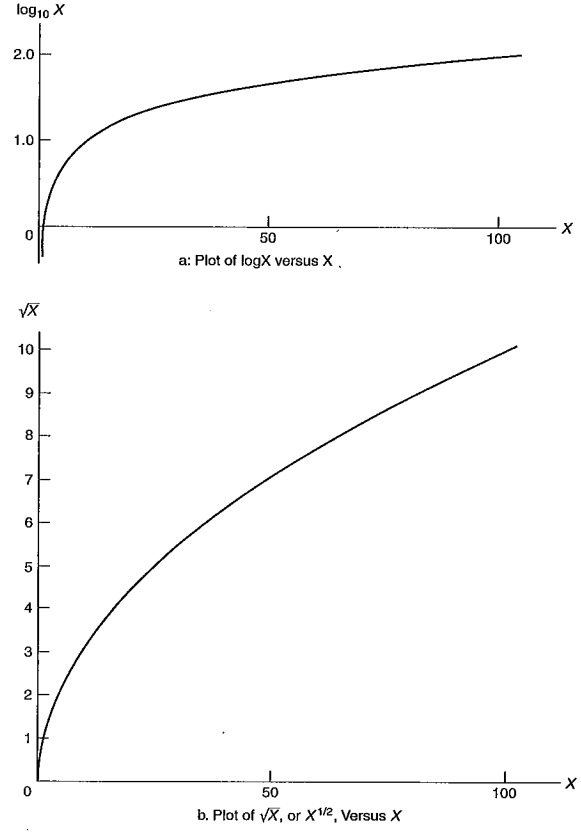


Figure 4.2: *Plots of Log and Square Root of X versus Variable X*

denoted by $\log_e$ or ln. The natural logarithm of X is the power to which e must be raised to produce X. There is a simple relationship between the natural and common logarithms, namely,

$$\log_e X = 2.3026 \log_{10} X$$
$$\log_{10} X = 0.4343 \log_e X$$

If we graph $\log_e X$ versus X, we would get a figure with the same shape as $\log_{10} X$, with the only difference being in the vertical scale; i.e., $\log_e X$ is larger than $\log_{10} X$ for $X > 1$ and smaller than $\log_{10} X$ for $X < 1$. The natural logarithm is used frequently in theoretical studies because of certain appealing mathematical properties.

Another class of transformations is known as **power transformations**. For example, the transformation X^2 (X raised to the power of 2) is used frequently

in statistical formulas such as computing the variance. The most commonly used power transformations are the square root (X raised to the power $\frac{1}{2}$) and the inverse $1/X$ (X raised to the power -1). Figure 4.2b shows a plot of the square root of X versus X. Note that this function is also not defined for negative values of X. Compared with taking the logarithm, taking the square root also progressively condenses the values of X as X increases. However, the degree of condensation is not as severe as in the case of logarithms. That is, the square root of X tapers off slower than $\log X$, as can be seen by comparing plots a and b in Figure 4.2.

Unlike $X^{\frac{1}{2}}$ and $\log X$, the function $1/X$ decreases with increasing X (see Figure 4.3a). To obtain an increasing function, you may use $-1/X$.

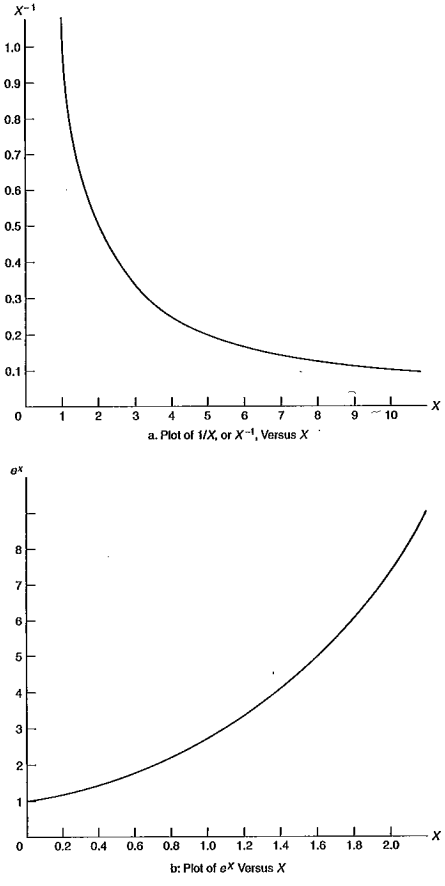


Figure 4.3: *Plots of* $1/X$ *and* e^X *versus the Variable* X

One way to characterize transformations is to note the power to which X is raised. We denote that power by p. For the square root transformation, $p = \frac{1}{2}$ and for the inverse transformation, $p = -1$. The logarithmic transformation can be thought of as corresponding to a value of $p = 0$ (Tukey, 1977). We can think of these three transformations as ranging from p values of -1 to 0 to $\frac{1}{2}$. The effects of these transformations in reducing a long-tailed distribution to the right are greater as the value of p decreases from 1 (the p value for no transformation) to $\frac{1}{2}$ to 0 to -1. Smaller values of p result in a greater degree of transformation for $p < 1$. Thus, p of -2 would result in a greater degree of transformation than a p of -1. The logarithmic transformation reduces a long tail more than a square root transformation (Figure 4.2a versus 4.2b) but not as much as the inverse transformation given in Figure 4.3a). The changes in the amount of transformation depending on the value of p provide the background for one method of choosing an appropriate transformation. In the next section, after a discussion of how to assess whether you have a normal distribution, we give several strategies to assist in choosing appropriate p values.

Finally, exponential functions are also sometimes used in data analysis. An **exponential function** of X may be thought of as the antilogarithm of X: for example, the antilogarithm of X to base 10 is 10 raised to the power X; similarly, the antilogarithm of X to base e is e raised to the power X. The exponential function e^x is illustrated in Figure 4.3b. The function 10^x has the same shape but increases faster than e^x. Both have the opposite effect of taking logarithms; i.e., they increasingly stretch the larger values of X. Additional discussion of these interrelationships can be found in Tukey (1977).

4.3 Selecting appropriate transformations

In the theoretical development of statistical methods some assumptions are usually made regarding the distribution of the variables being analyzed. Often the form of the distribution is specified. The most commonly assumed distribution for continuous observations is the **normal**, or **Gaussian**, distribution. Although the assumption is sometimes not crucial to the validity of the results, some check of normality is advisable in the early stages of analysis. For a review of the role of the normality assumption in the validity of statistical analyses, see Lumley *et al.* (2002). In this section, methods for assessing normality and for choosing a transformation to induce normality are presented.

Assessing normality using histograms

The left graph of Figure 4.4a illustrates the appearance of an ideal histogram, or density function, of normally distributed data. The values of the variable X are plotted on the horizontal axis. The range of X is partitioned into numerous intervals of equal length and the proportion of observations in each interval is

plotted on the vertical axis. The mean is in the center of the distribution and is equal to zero in this hypothetical histogram. The distribution is symmetric about the mean, that is, intervals equidistant from the mean have equal proportions of observations (or the same height in the histogram). If you place a mirror vertically at the mean of a symmetric histogram, the right side should be a mirror image of the left side. A distribution may be symmetric and still not be normal, but often distributions that are symmetric are close to normal. If the population distribution is normal, the sample histogram should resemble the famous symmetric bell-shaped Gaussian curve. For small sample sizes, the sample histogram may be irregular in appearance and the assessment of normality difficult.

Plots of various histograms and the appearance of their normal probability plot are given in Figure 4.4 and Figure 4.5.

All statistical packages described in Chapter 3 plot histograms for specific variables in the data set. The best fitting normal density function can also be superimposed on the histogram. Such graphs, produced by some packages, can enable you to make a crude judgment as to how well the normal distribution approximates the histogram.

Assessing symmetry using box plots

Symmetry of the empirical distribution can be assessed by examining certain percentiles (or quantiles). Quantiles are obtained by first ordering the N observations from smallest to largest. Using these ordered observations, the common rule for computing the quantile of the ith ordered observation is to compute $Q(i) = (i - 0.5)/N$. If N was five the third quantile would be $(3 - 0.5)/5 = .5$ or the median (see Cleveland, 1993). Note that there is a quantile for each observation. Percentiles are similar but they are computed to divide the sample into 100 equal parts. Quantiles or percentiles can be estimated from all six statistical packages. Usually plots are done of quantiles since in that case there is a point at each observation.

Suppose the observations for a variable such as income are ordered from smallest to largest. The person with an income at the 25th percentile would have an income such that 25% of the individuals have an income less than or equal to that person. This income is denoted as $P(25)$. Equivalently, the 0.25 quantile, often written as $Q(0.25)$, is the income that divides the cases into two groups where a fraction 0.25 of the cases has an income less than or equal to $Q(0.25)$ and a fraction 0.75 greater. Numerically $P(25) = Q(0.25)$. For further discussion on quantiles and how they are computed in statistical packages, see Frigge, Hoaglin and Iglewicz (1989).

Some quantiles are widely used. One such quantile is the sample median $Q(0.5)$. The median divides the observations into two equal parts. The median of ordered variables is either the middle observation if the number of observations N is odd, or the average of the middle two if N is even. Theoretically, for the normal distribution, the median equals the mean. In samples from normal distributions, they may not be precisely equal but they should not be far different. Two other widely used quantiles are $Q(0.25)$ and $Q(0.75)$. These are called "quartiles" since, along with the median, they divide the values of a variable into four equal parts.

If the distribution is symmetric, then the difference between the median and $Q(0.25)$ would equal the difference between $Q(0.75)$ and the median. This is displayed graphically in **box plots**. $Q(0.75)$ and $Q(0.25)$ are plotted at the top and the bottom of the box (or rectangle) and the median is denoted either by a line or a dot within the box. Lines (called **whiskers**) extend from the ends

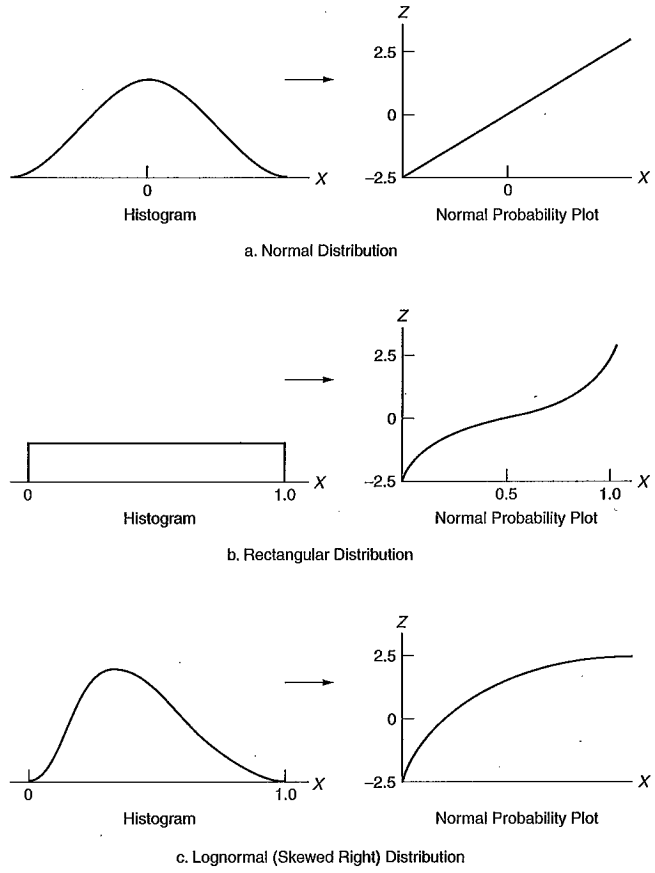
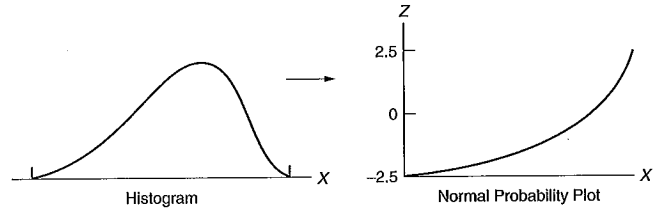
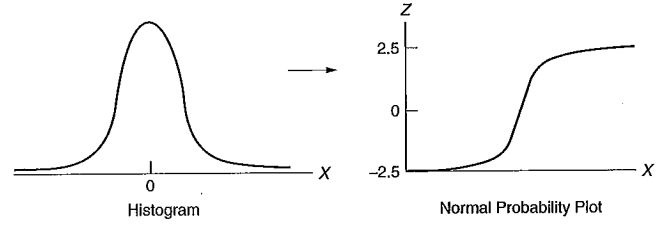


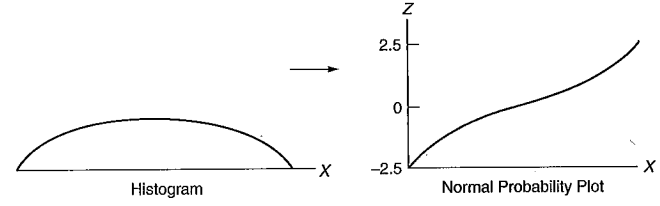
Figure 4.4: *Plots (a,b,c) of Common Histograms and the Resulting Normal Probability Plots from those Distributions*



d. Negative Lognormal (Skewed Left) Distribution



e. 1/Normal Distribution



f. High-Tail Distribution

Figure 4.5: *Plots (d,e,f) of Other Histograms and the Resulting Normal Probability Plots from those Distributions*

of the box out to what are called **adjacent values**. The numerical values of the adjacent values and quantiles are not precisely the same in all statistical packages (Frigge, Hoaglin and Iglewicz, 1989).

Usually the top adjacent value is the largest observation that is less than or equal to $Q(0.75)$ plus 1.5 ($Q(0.75) - Q(0.25)$). The bottom adjacent value is the smallest observation that is greater than or equal to $Q(0.25)$ minus $1.5(Q(0.75) - Q(0.25))$. Values that are beyond the adjacent value are sometimes examined to see if they are outliers. Symmetry can be assessed from box plots primarily by looking at the distances from the median to the top of the box and from the median to the bottom of the box. If these two lengths are decidedly unequal, then symmetry would be questionable. If the distribution is not symmetric, then it is not normal.

All six statistical packages can produce box plots.

Assessing normality using normal probability or normal quantile plots

Normal probability plots present an appealing option for checking for normality. One axis of the probability plot shows the values of X and the other shows expected values, Z, of X if its distribution were exactly normal. The computation of these expected Z values is discussed in Johnson and Wichern (2007). Equivalently, this graph is a plot of the **cumulative distribution** found in the data set, with the vertical axis adjusted to produce a straight line if the data followed an exact normal distribution. Thus if the data were from a normal distribution, the normal probability plot should approximate a straight line, as shown in the right-hand graph of Figure 4.4a. In this graph, values of the variable X are shown on the horizontal axis and values of the expected normal are shown on the vertical axis. Even when the data are normally distributed, if the sample size is small, the normal probability plot may not be perfectly straight, especially at the extremes. You should concentrate your attention on the middle 80% or 90% of the graph and see if that is approximately a straight line. Most investigators can visually assess whether or not a set of connected points follows a straight line, so this method of determining whether or not data are normally distributed is highly recommended.

In some packages the X and Z axes are interchanged. For your package, you should examine the axes to see which represents the data (X) and which represents the expected values (Z).

The remaining plots of Figure 4.4 and 4.5 illustrate other distributions with their corresponding normal probability plots. The purpose of these plots is to help you associate the appearance of the probability plot with the shape of the histogram or frequency distribution. One especially common departure from normality is illustrated in Figure 4.4c where the distribution is skewed to the right (has a longer tail on the right side). In this case, the data actually follow a **log-normal distribution**. Note that the curvature resembles a quarter of a circle (or an upside-down bowl) with the ends pointing downward. In this case we can create a new variable by taking the logarithm of X to either base 10 or e. A normal probability plot of $\log X$ will follow a straight line.

If the axes in Figure 4.4c were interchanged so X was on the vertical axis, then the bowl would be right side up with the ends pointing upward. It would look like Figure 4.5d.

Figures 4.4b and 4.5f show distributions that have higher tails than a normal distribution (more extreme observations). The corresponding normal probability plot is an inverted S. Figure 4.5e illustrates a distribution where the tails have fewer observations than a normal does. It was in fact obtained by taking the inverse of normally distributed data. The resulting normal probability plot is S-shaped. If the inverse of the data $(1/X)$ were plotted, then a straight line

normal probability plot would be obtained. If X were plotted on the vertical axis then Figure 4.5e would illustrate a heavy tail distribution and Figures 4.4b and 4.5f would illustrate distributions that have fewer expected observations than in the tails of a normal distribution.

The advantage of using the normal probability plot is that it not only tells you if the data are approximately normally distributed, but it also offers insight into how the data are distributed if they are not normally distributed. This insight is helpful in deciding what transformation to use to induce normality.

Many software programs also produce theoretical quantile–quantile plots. Here a theoretical distribution such as a normal is plotted against the empirical quantiles of the data. When a normal distribution is used, they are often called **normal quantile** plots. These plots are interpreted similarly to normal probability plots with the difference being that quantile plots emphasize the tails more than the center of the distribution. Normal quantile plots may have the variable X plotted on the vertical axis in some packages (see S-PLUS). For further information on their interpretation, see Gan, Koehler and Thompson (1991).

Normal probability or normal quantile plots are available also in all six packages.

Selecting a transformation to induce normality

As we noted in the previous discussion, transforming the data can sometimes produce an approximate normal distribution. In some cases the appropriate transformation is known. Thus, for example, the square root transformation is used with Poisson-distributed variables. Such variables represent counts of events occurring randomly in space or time with a small probability such that the occurrence of one event does not affect another. Examples include the number of cells per unit area on a slide of biological material, the number of incoming phone calls per second in a telephone exchange, and counts of radioactive material per unit of time. Similarly, the logarithmic transformation has been used on variables such as household income, the time elapsing before a specified number of Poisson events have occurred, systolic blood pressure of older individuals, and many other variables with long tails to the right. Finally, the inverse transformation has been used frequently on the length of time it takes an individual or an animal to complete a certain task. Tukey (1977) gives numerous examples of variables and appropriate transformations on them to produce approximate normality. Tukey (1977), Box and Cox (1964), Draper and Hunter (1969), and Bickel and Doksum (1981) discuss some systematic ways to find an appropriate transformation. A table of common transformations is included in van Belle *et al.* (2004).

Another strategy for deciding what transformation to use is to progress up or down the values of p depending upon the shape of the normal probability

plot. If the plot looks like Figure 4.4c, then a value of p less than 1 is tried. Suppose $p = \frac{1}{2}$ is tried. If this is not sufficient and the normal probability plot is still curved downward at the ends, then $p = 0$, i.e., the logarithmic transformation, can be tried. If the plot appears to be almost correct but a little more transformation is needed, a positive constant can be subtracted from each X prior to the transformation. If the transformation appears to be a little too much, a positive constant can be added. Thus, various transformations of the form $(X + C)^p$ are tried until the normal probability plot is as straight as possible. An example of the use of this type of transformation occurs when considering systolic blood pressure (SBP) for older adult males. It has been found that a transformation such as log(SBP – 75) results in data that appear to be normally distributed.

The investigator decreases or increases the value of p until the resulting observations appear to have close to a normal distribution. Note that this does not mean that theoretically you have a normal distribution since you are only working with a single sample. Particularly if the sample is small, the transformation that is best for your sample may not be the one that is best for the entire population. For this reason, most investigators tend to round off the numerical values of p. For example, if they found that $p = 0.46$ worked best with their data, they might actually use $p = 0.5$ or the square root transformation.

All the statistical packages discussed in Chapter 3 allow the user to perform a large variety of transformations. Stata has an option, called ladder of powers, that automatically produces power transformations for various values of p.

Hines and O'Hara Hines (1987) developed a method for reducing the number of iterations needed by providing a graph which produces a suggested value of p from information available in most statistical packages. Using their graph one can directly estimate a reasonable value of p.

Their method is based on the use of quantiles. What are needed for their method are the values of the median and a pair of symmetric quantiles. As noted earlier, the normal distribution is symmetric about the median. The difference between the median and a lower quantile (say $Q(0.2)$) should equal the difference between the upper symmetric quantile (say $Q(0.8)$) and the median if the data are normally distributed. By choosing a value of p that results in those differences being equal, one is choosing a transformation that tends to make the data approximately normally distributed. Using Figure 4.6 (kindly supplied by W. G. S. Hines and R. J. O'Hara Hines), one plots the ratio of the lower quantile of X to the median on the vertical axis and the ratio of the median to the upper quantile on the horizontal axis for at least one set of symmetric quantiles. The resulting p is read off the closest curve.

For example, in the depression data set given in Table 3.4 a variable called INCOME is listed. This is family income in thousands of dollars. The median is 15, $Q(0.25) = 9$, and $Q(0.75) = 28$. Hence the median is not halfway between the two quantiles (which in this case are the quartiles), indicating a nonsymmetric distribution with the long tail to the right. If we plot $9/15 = 0.60$ on

the vertical axis and $15/28 = 0.54$ on the horizontal axis of Figure 4.3, we get a value of p of approximately $-\frac{1}{3}$. Since $-\frac{1}{3}$ lies between 0 and $-\frac{1}{2}$, trying first the log transformation seems reasonable. The median for log(INCOME) is 1.18, $Q(0.25) = 0.95$, and $Q(0.75) = 1.45$. The median $- 0.95 = 0.23$ and 1.45 – median $= 0.27$, so from the quartiles it appears that the data are still slightly skewed to the right. Additional quantiles could be estimated to better approximate p from Figure 4.6. When we obtained an estimate of the skewness of log(INCOME), it was negative, indicating a long tail to the left. The reason for the contradiction is that seven respondents had an income of only 2 ($2000 per year) and these low incomes had a large effect on the estimate of the skewness but less effect on the quartiles. The skewness statistic is sensitive to extreme values.

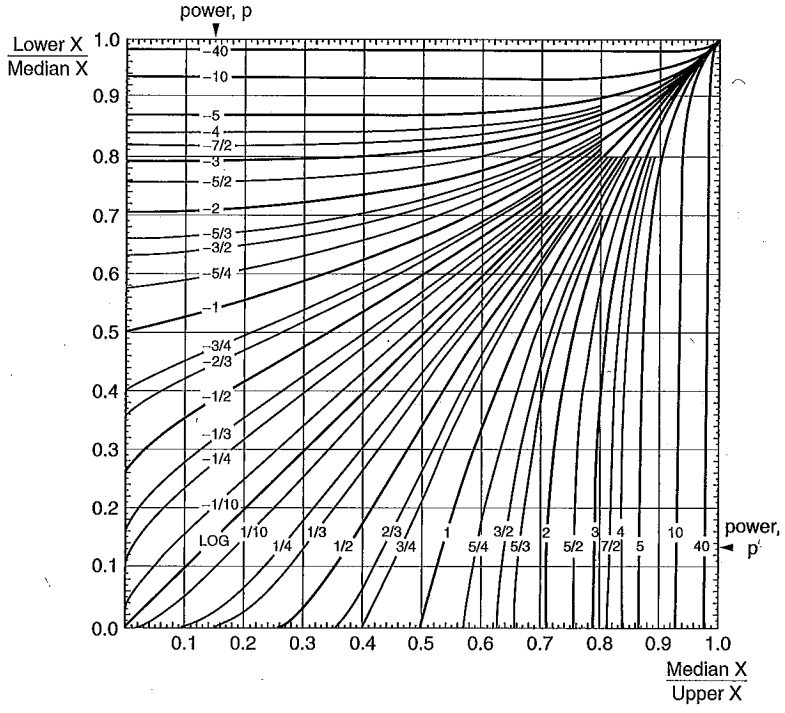


Figure 4.6: *Determination of the Value of p in the Power Transformation to Produce Approximate Normality*

Stata, through the command "lnskew0," considers all transformations of the form $\ln(X - k)$ and finds the value of k which makes the skewness approximately equal to 0. For the variable INCOME in the depression data set, the value of k chosen by the program is $k = -2.01415$. For the transformed

variable ln(INCOME + 2.01415) the value of the skewness is -0.00016 compared with 1.2168 for INCOME. Note that adding a constant reduces the effect of taking the log. Examining the histograms with the normal density imposed upon them shows that the transformed variable fits a normal distribution much better than the original data. There does not appear to be much difference between the log transformation suggested by Figure 4.3 and the transformation obtained from Stata when the histograms are examined.

It should be noted that not every distribution can be transformed to a normal distribution. For example, variable 29 in the depression data set is the sum of the results from the 20-item depression scale. The mode (the most commonly occurring score) is at zero, thus making it virtually impossible to transform these CESD scores to a normal distribution. However, if a distribution has a single mode in the interior of the distribution, then it is usually not too difficult to find a distribution that appears at least closer to normal than the original one. Ideally, you should search for a transformation that has the proper scientific meaning in the context of your data.

Statistical tests for normality

It is also possible to do formal tests for normality. These tests can be done to see if the null hypothesis of a normal distribution is rejected and hence whether a transformation should be considered. They can also be used after transformations to assist in assessing the effect of the transformation. A commonly used procedure is the Shapiro–Wilk W statistic (Mickey, Dunn and Clark, 2009). This test has been shown to have good power against a wide range of non-normal distributions (see D'Agostino, Belanger and D'Agostino (1990) for a discussion of this test and others). Several of the statistical programs give the value of W and the associated p value for testing that the data came from a normal distribution. If the value of p is small, then the data may not be considered normally distributed. For example, we used the command "swilk" in Stata to perform the Shapiro–Wilk test on INCOME and ln(INCOME + 2.01415) from the depression data set. The p values for the two variables were 0.00000 and 0.21347, respectively. These results indicate that INCOME has a distribution that is significantly different from normal, while the transformed data fit a normal distribution reasonably well. A more practical transformation for data analysis is ln(INCOME + 2). This transformed variable is also fitted well with a normal distribution ($p = 0.217$).

Some programs also compute the Kolmogorov–Smirnov D statistic and approximate p values. Here the null hypothesis is rejected for large D values and small p values. van Belle *et al.* (2004) discuss the characteristics of this test. It is also possible to perform a chi-square goodness of fit test. Note that both the Kolmogorov–Smirnov D test and the chi-square test have poor power properties. In effect, if you use these tests you will tend to reject the null hypothesis

when your sample size is large and accept it when your sample size is small. As one statistician put it, they are to some extent a complicated way of determining your sample size.

Another way of testing for normality is to examine the **skewness** of your data. Skewness is a measure of how nonsymmetric a distribution is. If the data are symmetrically or normally distributed, the computed skewness will be close to zero. The numerical value of the ratio of the skewness to its standard error can be compared with normal Z tables, and symmetry, and hence normality, would be rejected if the absolute value of the ratio is large. Positive skewness indicates that the distribution has a long tail to the right and probably looks like Figure 4.4c. (Note that it is important to remove outliers or erroneous observations prior to performing this test.) D'Agostino, Belanger and D'Agostino (1990) give an approximate formula for the test of skewness for sample sizes > 8. Skewness is computed in all six packages. STATISTICA provides estimates of the standard error that can be used with large sample sizes. Stata computes the test of D'Agostino, Belanger and D'Agostino (1990) via the sktest command. In general, a distribution with skewness greater than one will appear to be noticeably skewed unless the sample size is very small.

Very little is known about what significance levels α should be chosen to compare with the p value obtained from formal tests of normality. So the sense of increased preciseness gained by performing a formal test over examining a plot is somewhat of an illusion. From the normal probability plot you can both decide whether the data are normally distributed and get a suggestion about what transformation to use.

Assessing the need for a transformation

In general, transformations are more effective in inducing normality when the standard deviation of the untransformed variable is large relative to the mean. If the standard deviation divided by the mean is less than $\frac{1}{4}$, then the transformation may not be necessary. Alternatively, if the largest observation divided by the smallest observation is less than 2, then the data are likely not to be sufficiently variable for the transformation to have a decisive effect on the results (Hoaglin, Mosteller and Tukey, 1985). These rules are not meaningful for data without a natural zero (interval data but not ratio data by Stevens's classification). For example, the above rules will result in different values if temperature is measured in Fahrenheit or Celsius. For variables without a natural zero, it is often possible to estimate a natural minimum value below which observations are seldom if ever found. For such data, the rule of thumb is if

$$\frac{\text{Largest value} - \text{natural minimum}}{\text{smallest observation} - \text{natural minimum}} < 2$$

then a transformation is not likely to be useful.

Other investigators examine how far the normal probability plot is from a straight line. If the amount of curvature is slight, they would not bother to transform the data.

Note that the usefulness of transformations is difficult to evaluate when the sample sizes are small or if numerous outliers are present. In deciding whether to make a transformation, you may wish to perform the analysis with and without the proposed transformation. Examining the results will frequently convince you that the conclusions are not altered after making the transformation. In this case it is preferable to present the results in terms of the most easily interpretable units. And it is often helpful to conform to the customs of the particular field of investigation.

Sometimes, transformations are made to simplify later analyses rather than to approximate normal distributions. For example, it is known that FEV1 (forced expiratory volume in 1 second) and FVC (forced vital capacity) decrease in adults as they grow older. (See Section 1.3 for a discussion of these variables.) Some researchers will take the ratio FEV1/FVC and work with it because this ratio is less dependent on age. Using a variable that is independent of age can make analyses of a sample including adults of varying ages much simpler. In a practical sense, then, the researcher can use the transformation capabilities of the computer program packages to create new variables to be added to the set of initial variables rather than only modify and replace them.

If transformations alter the results, then you should select the transformation that makes the data conform as much as possible to the assumptions. If a particular transformation is selected, then all analyses should be performed on the transformed data, and the results should be presented in terms of the transformed values. Inferences and statements of results should reflect this fact.

4.4 Assessing independence

Measurements on two or more variables collected from the **same** individual are not expected to be, nor are they assumed to be, independent of each other. On the other hand, independence of observations collected from **different** individuals or items is an assumption made in the theoretical derivation of most multivariate statistical analyses. This assumption is crucial to the validity of the results, and violating it may result in erroneous conclusions. Unfortunately, little is published about the quantitative effects of various degrees of nonindependence. Also, few practical methods exist for checking whether the assumption is valid.

In situations where the observations are collected from people, it is frequently safe to assume independence of observations collected from different individuals. Dependence could exist if a factor or factors exist to affect all of the individuals in a similar manner with respect to the variables being mea-

sured. For example, political attitudes of adult members of the same household cannot be expected to be independent. Inferences that assume independence are not valid when drawn from such responses. Similarly, biological data on twins or siblings are usually not independent.

Data collected in the form of a sequence either in time or space can also be dependent. For example, observations of temperature on successive days are likely to be dependent. In those cases it is useful to plot the data in the appropriate sequence and search for trends or changes over time. Some programs allow you to plot an outcome variable against the program's own ID variable. Other programs do not; so the safe procedure is to always type in a variable or variables that represent the order in which the observations are taken and any other factor that you think might result in lack of independence, such as location or time. Also, if more complex sampling methods are used, then information on clusters or strata should be entered.

Figure 4.7 presents a series of plots that illustrate typical outcomes. Values of the outcome variable being considered appear on the vertical axis and values of order of time, location, or observation identification number (ID) appear on the horizontal axis. If the plot resembles that shown in Figure 4.7a, little reason exists for suspecting lack of independence or nonrandomness. The data in that plot were, in fact, obtained as a sequence of random standard normal numbers. In contrast, Figure 4.7b shows data that exhibit a positive trend. Figure 4.7c is an example of a temporary shift in the level of the observations, followed by a return to the initial level. This result may occur, for example, in laboratory data when equipment is temporarily out of calibration or when a substitute technician, following different procedures, temporarily performs the work. Finally, a common situation occurring in some business or industrial data is the existence of seasonal cycles, as shown in Figure 4.7d.

Such plots or scatter diagrams are available in all six statistical packages mentioned in this book and are widely available elsewhere. If a two-dimensional display of location is desired with the outcome variable under consideration on the vertical axis, then this can be displayed in three dimensions in some packages.

Some formal tests for the randomness of a sequence of observations are given in Brownlee (1965). One such test (the Durbin-Watson test) is presented in Chapter 6 of this book in the context of regression and correlation analysis. If you are dealing with series of observations you may also wish to study the area of forecasting and time series analysis. Some books on this subject are referenced in Chapter 6.

4.5 Summary

In this chapter we emphasized methods for determining if the data were normally distributed and for finding suitable transformations to make the data

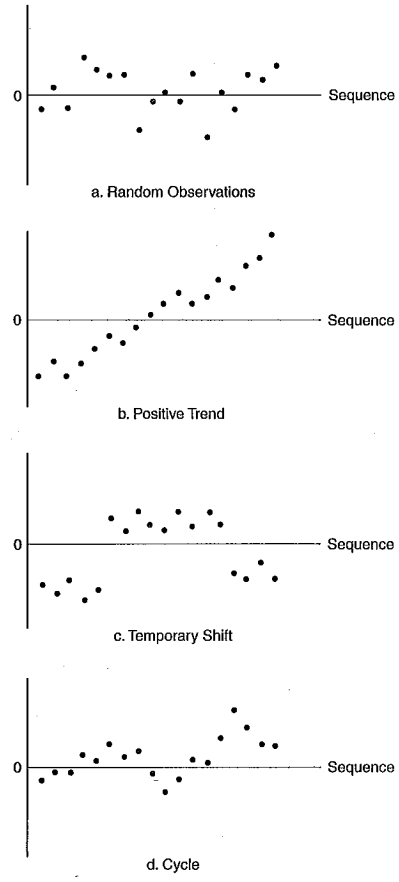


Figure 4.7: *Graphs of Hypothetical Data Sequences Showing Lack of Independence*

closer to a normal distribution. We also discussed methods for checking whether the data were independent.

No special order is best for all situations in data screening, However, most investigators would delete the obvious outliers first. They may then check the assumption of normality, attempt some transformations, and recheck the data for outliers. Other combinations of data-screening activities are usually dictated by the problem.

We wish to emphasize that data screening can be the most time-consuming and costly portion of data analysis. Investigators should not underestimate this aspect. If the data are not screened properly, much of the analysis may have to be repeated, resulting in an unnecessary waste of time and resources. Once the data have been carefully screened, the investigator will then be in a position to

select the appropriate analysis to answer specific questions. In Chapter 5 we present a guide to the selection of the appropriate data analysis.

4.6 Problems

4.1 Using the depression data set described in Tables 3.3 and 3.4, create a variable equal to the negative of one divided by the cubic root of income. Display a normal probability plot of the new variable.

4.2 Take the logarithm to base 10 of the income variable in the depression data set. Compare the histogram of income with the histogram of log(INCOME). Also, compare the normal probability plots of income and log(INCOME).

4.3 Repeat Problems 4.1 and 4.2, taking the square root of income.

4.4 Generate a set of 100 random normal deviates by using a computer program package. Display a histogram and normal probability plot of these values. Square these numbers by using transformations. Compare histograms and normal probability plots of the logarithms and the square roots of the set of squared normal deviates.

4.5 Use the two sets of 100 random numbers from Problem 4.4. Display boxplots of these two sets of values and state which of the three graphical methods (histograms, normal probability plots, and boxplots) are most useful in assessing normality.

4.6 Take the logarithm of the CESD score plus 1 and compare the histograms of CESD and log(CESD + 1). (A small constant must be added to CESD because CESD can be zero.)

4.7 Obtain normal probability plots of mothers' and fathers' weights from the lung function data set described in Appendix A. Discuss whether or not you consider weight to be normally distributed in the population from which this sample is taken.

4.8 The accompanying data are from the New York Stock Exchange Composite Index for the period August 9 through September 17, 1982. Run a program to plot these data in order to assess the lack of independence of successive observations. (Note that the data cover five days per week, except for a holiday on Monday, September 6). The daily volume is the number of transactions given in millions of shares. This time period saw an unusually rapid rise in stock prices (especially for August), coming after a protracted falling market. Compare this time series with prices for the current year.

Day	Month	Index	Volume	Day	Month	Index	Volume
9	Aug.	59.3	63	1	Sept.	67.9	98
10	Aug.	59.1	63	2	Sept.	69.0	87
11	Aug.	60.0	59	3	Sept.	70.3	150
12	Aug.	58.8	59	6	Sept.	–	–
13	Aug.	59.5	53	7	Sept.	69.6	81
16	Aug.	59.8	66	8	Sept.	70.0	91
17	Aug.	62.4	106	9	Sept.	70.0	87
18	Aug.	62.3	150	10	Sept.	69.4	82
19	Aug.	62.6	93	13	Sept.	70.0	71
20	Aug.	64.6	113	14	Sept.	70.6	98
23	Aug.	66.4	129	15	Sept.	71.2	83
24	Aug.	66.1	143	16	Sept.	71.0	93
25	Aug.	67.4	123	17	Sept.	70.4	77
26	Aug.	68.0	160				
27	Aug.	67.2	87				
30	Aug.	67.5	70				
31	Aug.	68.5	100				

4.9 Generate ten random normal deviates. Display a probability plot of these data. Suppose you didn't know the origin of these data. Would you conclude they were normally distributed? What is your conclusion based on the Shapiro–Wilk test? Do ten observations provide sufficient information to check normality?

4.10 Obtain a normal probability plot of the index given in Problem 4.8. Suppose that you had been ignorant of the lack of independence of these data and had treated them as if they were independent samples. Assess whether they are normally distributed.

4.11 Repeat problem 4.7 with weights expressed in ounces instead of pounds. How will your conclusions change? Obtain normal probability plots of the logarithm of mothers' weights expressed in pounds and then in ounces, and compare.

4.12 From the variables ACUTEILL and BEDDAYS described in Table 3.3, create a single variable that takes on the value 1 if the person has been both bedridden and acutely ill in the last two months and that takes on the value 0 otherwise.

4.13 Using the Parental HIV data set (see Appendix A), plot a histogram, boxplot, and a normal probability plot for the variable AGESMOKE. This variable is the age in years when the respondent started smoking. If the respondent did not start smoking, AGESMOKE was assigned a value of zero. Decide what to do about the zero values and if a transformation should be

used for this variable if the assumption of normality is made when it is used in a statistical analysis.

4.14 Using the Parental HIV data calculate an overall Brief Symptom Inventory (BSI) score for each adolescent (see the codebook for details). Log-transform the BSI score. Obtain a normal probability plot for the log-transformed variables. Does the log-transformed variable seem to be normally distributed? As you might notice, the numbers of adolescents with a missing value on the overall BSI score and the log-transformed BSI score are different. Why is this the case? Could this influence our conclusions regarding the normality of the transformed variable? How could this be avoided?

4.15 Using the lung cancer data described in Appendix A, examine the distribution of the variable *days* separately for those who died (death=1) and for those who did not (death=0). Plot a normal probability plot, a histogram, and a boxplot for each. Use the methods described in this chapter to choose appropriate transformations to induce approximate normality. Are the chosen transformations the same for the two groups? Discuss the results.

Chapter 5

Selecting appropriate analyses

5.1 Which analyses to perform?

When you have a data set and wish to analyze it using the computer, obvious questions that arise are "What descriptive measures should be used to examine the data?" and "What statistical analyses should be performed?" Section 5.2 explains why sometimes these are confusing questions, particularly to investigators who lack experience in real-life data analysis. Section 5.3 presents measures that are useful in describing data. Usually, after the data have been "cleaned" and any needed transformations made, the next step is to obtain a set of descriptive statistics and graphs. It is also useful to obtain the number of missing values for each variable. The suggested descriptive statistics and graphs are guided by Stevens's classification system introduced in Chapter 2. Section 5.4 presents a table that summarizes suggested multivariate analyses and indicates in which chapters of this book they can be found. Readers with experience in data analysis may be familiar with all or parts of this chapter.

5.2 Why selection is often difficult

There are two reasons why deciding what descriptive measures or analyses to perform and report is often difficult for an investigator with real-life data. First, in statistics textbooks, statistical methods are presented in a logical order from the viewpoint of learning statistics but not from the viewpoint of doing data analysis by using statistics. Most texts are either mathematical statistics texts, or are imitations of them with the mathematics simplified or left out. Also, when learning statistics for the first time, the student often finds mastering the techniques themselves tough enough without worrying about how to use them in the future. The second reason is that real-life data often contain mixtures of types of data, which makes the choice of analysis somewhat arbitrary. Two trained statisticians presented with the same set of data will often opt for different ways of analyzing the set, depending on what assumptions they are willing to take into account in the interpretation of the analysis.

Acquiring a sense of when it is safe to ignore assumptions is difficult both