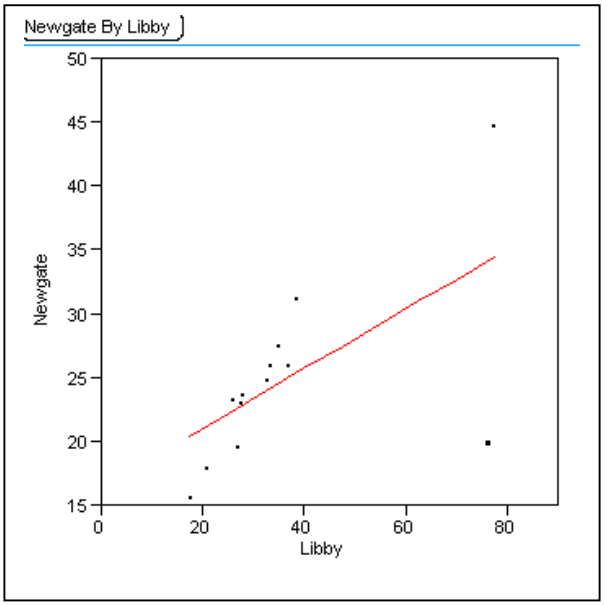
Ye Olde
Assessment of the Linear Fit

Looking at the scatter plot and the fit line seems like such an obvious step to beginning data analysts; in this case the beginning data analysts are absolutely correct. The trouble comes with a little experience when the no longer neophyte data analyst begins to understand and depend on numeric computer output and thus pays less attention to the original data. A good rule to follow: check the scatter plot first, then look at the numeric output. A more than passing acquaintance with the plot of the data will better inform your analysis of the numeric information.

What should we look for in a regression plot? The short answer is, anything unusual. It is possible for individual isolated points to pique our interest; or perhaps a cluster or a pattern of points may arouse our attention. In either situation the potential effect of unusual situations needs to be evaluated.

Isolated Points as Symptoms

Consider the scatterplot below, relating the stream flow at two locations on the Kootenai River: Newgate, British Columbia and Libby, Montana, further downstream. As part of planning for a dam the past history of the stream flow was gathered. (U.S. Dept of Interior, 1947)

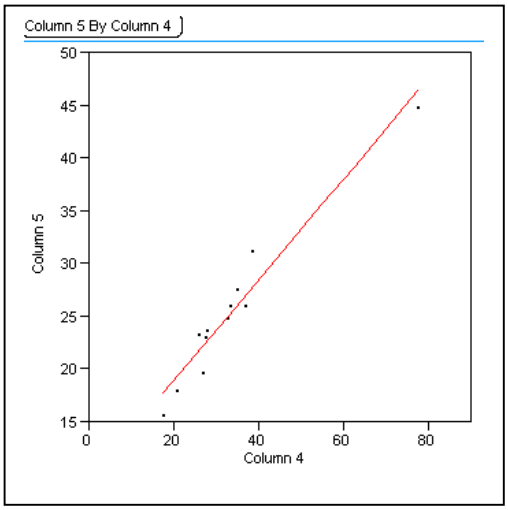


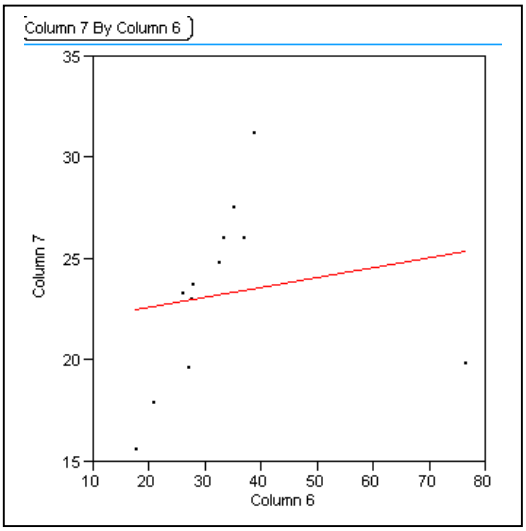
There is a cluster of points in this scatterplot, and then two points which are isolated. The idea of isolated points is reminiscent of the idea of "outliers" in univariate data analysis. There is no generally accepted definition of an outlier in the setting of a scatter plot, but the terminology in the sense of a data element off by itself certainly applies. A point may be "outlying" because of an unusually high or low horizontal or vertical coordinate, or may have an extreme value for both coordinates. A data point away from the rest of the pattern could represent a mis-entry of data, an unusual circumstance (in this case, a year-long drought), or may be a signal to the data analyst of a more complex relation than heretofore suspected.

When a best fit line, $\hat{y} = a + bx$, is generated from the data, outliers represent another potential problem: they may exert more than their share of influence on the regression algorithm. A principle in simple linear regression is that each point contributes "equally" to the construction of the best-fit line. But just as an outlier in a univariate set of data can produce a mean not representative of the data as a whole, an extreme point can produce a best fit line which is not representative of the underlying relation between the two variables. A point that by itself has a significant impact on the best fit line is

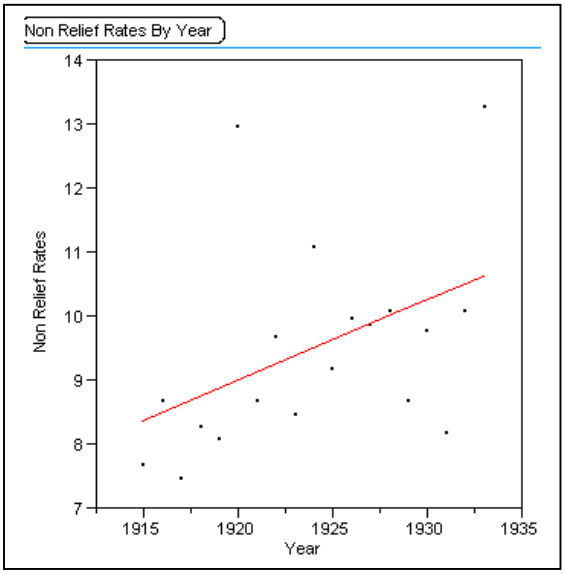
called an "influential" point. Influential points are located by a fairly easy method: exclude them from the data set and recalculate the best-fit line. If the intercept and/or slope of the regression line is significantly changed, the point is by definition influential. Advanced books on regression discuss numerical methods for assessing the influence of a point. (Draper & Smith, 1998; Myers, 1986)

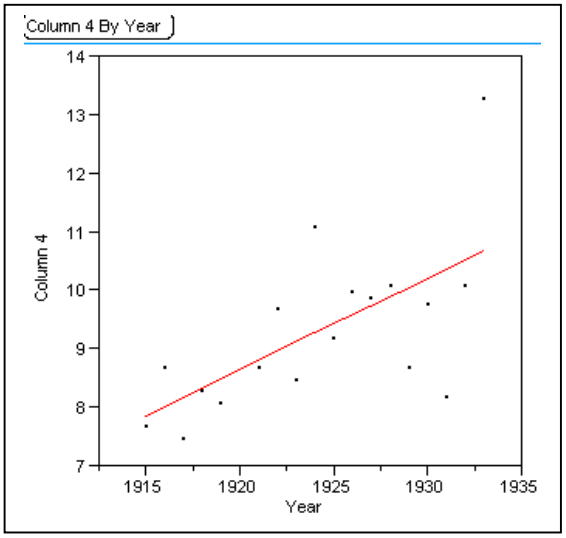
The influence of a point is graphically (and dramatically !) presented below. The two outlying points in the Kootenai River data are (77.6, 44.9) and (76.6, 19.9). For the two plots below these points have been deleted one at a time. Notice that the deletion of either point has a distinct effect on both the slope and intercept of the best fit line.





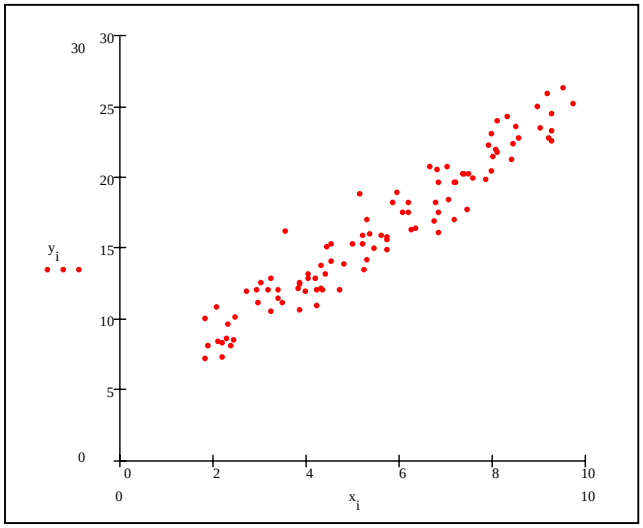
It is also possible for a point to exert its influence on the intercept while having little effect on the slope of a best-fit line. The scatterplot below shows the marriage rate of persons not on public relief rolls, sampled from five rural counties in North Carolina, for the years 1915-1933. (Hamilton, 1936-) Deleting the point for the year 1920 has a distinct effect on the intercept of the line, but not very much change in the slope.





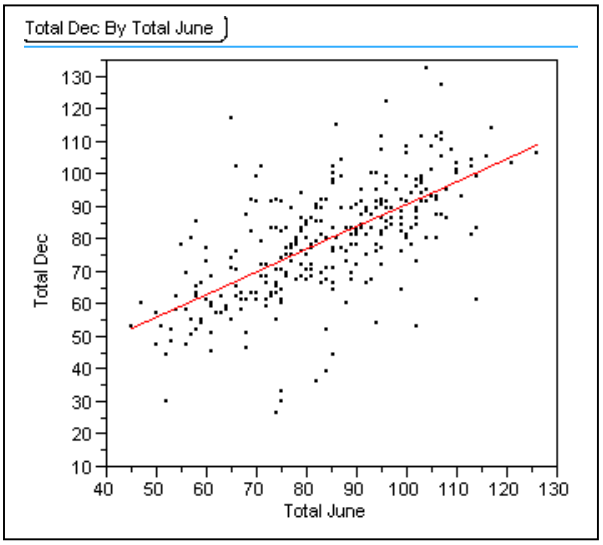
Patterns of points as Symptoms

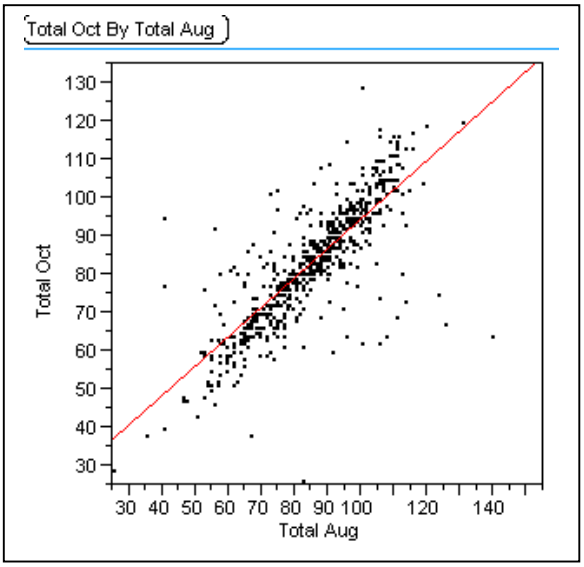
When a line is fit to data, it is expected that the points generally line up in a parallelogram with the least squares line dividing the parallelogram into two halves, upper and lower. The points should tend to be denser near the best fit line than they are farther away. Any deviation from this pattern is of interest when assessing the fit of a line. There are many ways that a scatterplot could differ from the standard parallelogram shape. One possibility is the points will follow a curve, rather than a straight line. Another possibility is the points will arrange themselves in a seriously non-parallelogram form. Or there may be a parallelogram pattern to the points, but a few points seem to be misplaced.



Below are four scatter plots with their least squares best fit lines. Three of the scatter plots depart from the expected parallelogram form. The first two scatter plots are historical in nature, dating from the War Between the States. These data were gathered from Confederate muster rolls for the latter half of 1861. (C. Olsen, 1995) Early in the conflict prospective soldiers would gather by geographic region at a county seat or other designated location and volunteer. When there were enough men to form a company --

by law, at least 64 men plus officers -- they would depart for service in the army. Individuals who volunteered later than the initial company formation would join the company formed in their geographic area; thus, differences in company sizes over a six month period would reflect recruitment not only at the beginning of the War but during the first six months. The investigator in this study was interested in the sizes of these companies.

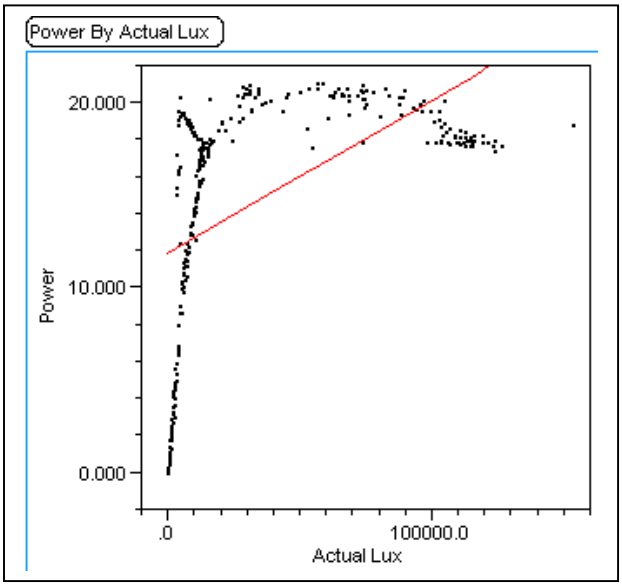


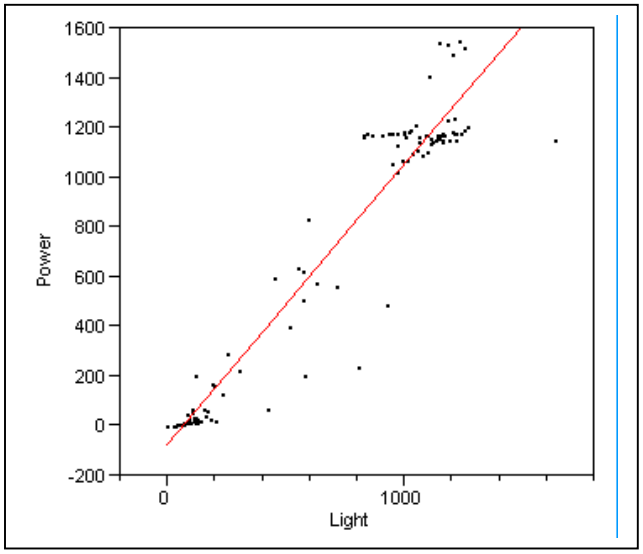


The plot of Company sizes in December vs. Company sizes in June is what might be thought typical for a linear regression. The points of the plot follow the best fit line, and their scatter about the line is generally consistent with a parallelogram sort of shape. Consider now the October vs. August plot. This plot shows a cluster of points with a very tight fit to the best fit line, some points that seem to buck the trend. A significant number of points are higher than one might expect for low August company sizes, and there are also points which are lower than would be predicted for companies with higher August totals. These wayward points are of interest precisely because they are wayward, making the scatter plot different from the "classic" parallelogram plot.

The next two plots are from a study of material for solar panels. The investigators (A. Olsen & J. Veenstra, 1999) were interested in the relationship between the amount of sunlight and the power generated by the material. The most obvious feature of the plot on the left below is its non-linearity! There certainly appears to be a relation between the two variables, but it does not appear to be a linear relationship. It is of interest also to note two anomalies. There seems to be a distinct pattern of points in the upper left part of the plot that branch off in their own direction. Another problem for the investigators would seem to be the decrease in power as the amount of sunlight increases when the value of the explanatory variable exceeds about 60,000. Neither the non-linearity nor the other interesting features would be detected by studying the numeric computer output only.

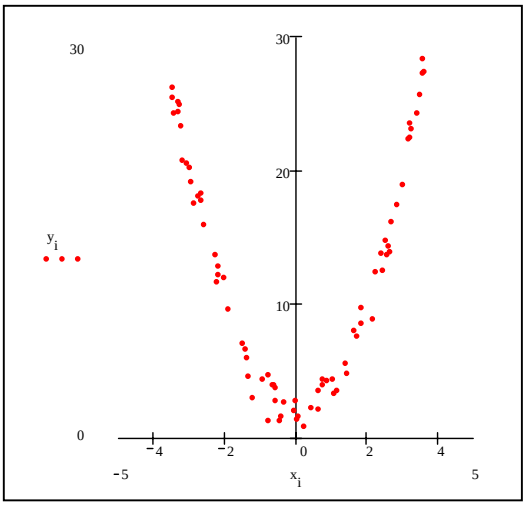
The plot on the right perfectly illustrates why it is crucial to look at the scatter plot of data. Again, to belabor the point, had the investigators only looked at the numeric output they would see a very high r^2 and might be misled into believing the line to be a good fit to the points. However, the scatter plot reveals the data in two clumps, perhaps the result of a sudden appearance of clouds on a very sunny day.





The examples presented above should send a clear message to the data analyst: don't just generate a scatterplot of data, <u>study</u> it. No numeric summaries can approach the richness of information contained in the plot.

Checking the residuals



When assessing the fit of a line to data, evaluating the pattern of residuals is just as important as attending to their lengths. Small residuals result in small sums of squares of residuals which, when compared to the total sum of squares, lead to a large coefficient of determination. However, the size of r^2 tells only part of the story. Two variables could be very clearly related but have a low (or even zero) r^2, such as the parabolic scatter plot below. And, as we saw with the clumped data above, a high correlation is possible when no linear relation has yet been demonstrated. When a straight line is fit to data we are there is an

assumption -- naturally enough -- that the underlying relationship between the variables is linear. The mathematical formulae that result in the best fit line produce a line which is hopefully a good representation of the assumed underlying relation between the variables. But just like carpenters' tools, statistical tools need to be treated with caution. We should not be surprised at failing if we try to use a hammer for tightening bolts; analogously, using linear regression to analyze data where the underlying relation is not linear may also give us unexpected and misleading results.

Diagnosing problems in linear regression can be accomplished effectively using elementary graphic techniques. The data analyst proceeds much as the family doctor: she searches for symptoms of problems. In linear regression analysis checking for symptoms involves examination of the original scatter plot, seeking out any unduly influential points, and especially studying the residuals. The table below presents the characteristics of the residuals for good fits and fits that may be less so.

Table X

Residuals from a "good" linear fit	**Residuals from a "problem" linear fit**
The residuals are approximately normally distributed AND	The residuals are <u>not</u> even approximately normally distributed AND/OR
the residuals are equally variable for all values of the explanatory variable AND	the residuals are <u>not</u> equally variable for all values of the explanatory variable AND/OR
the residuals don't exhibit any discernible "pattern."	the residuals do exhibit a discernible "pattern."

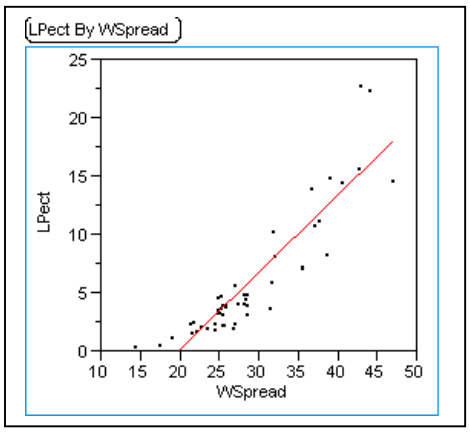
For our discussion of methods for evaluating the symptoms we will use some data describing physical characteristics of 49 species of flying birds from France (Greenwalt, 1962.) The data are easy to understand, and one need not be an expert ornithologist to interpret them. Here are some terms you will see in the discussion to follow. The wing length and wing span are both linear measures of wing size. The pectoral muscles for the bird are the muscles that power the birds' flight; the large pectoral muscle powers the downbeat of the wing stroke, the small pectoral muscle powers the upbeat.

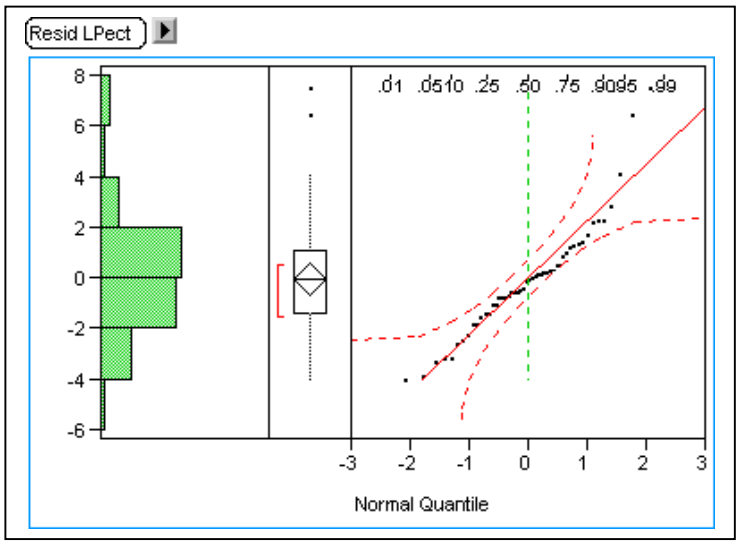
Detecting non-normality of residuals: the histogram and normal probability plot

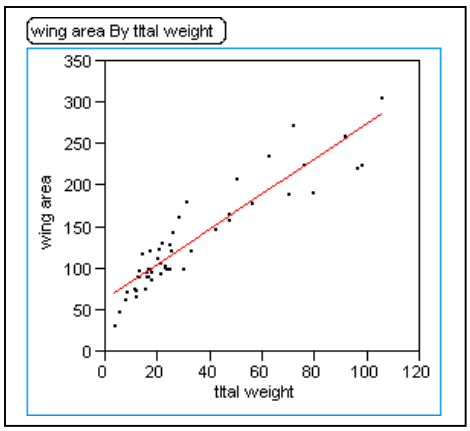
Some of the most frequently used statistical methods are valid only when a sample, $x_1, x_2, \ldots, x_n$ is from a population distribution that is at least approximately normal. Because of this statisticians have developed methods for detecting non-normality and thus, detecting the potential invalidity of a statistical method under consideration. We will use some of these techniques to assess the normality of residuals.

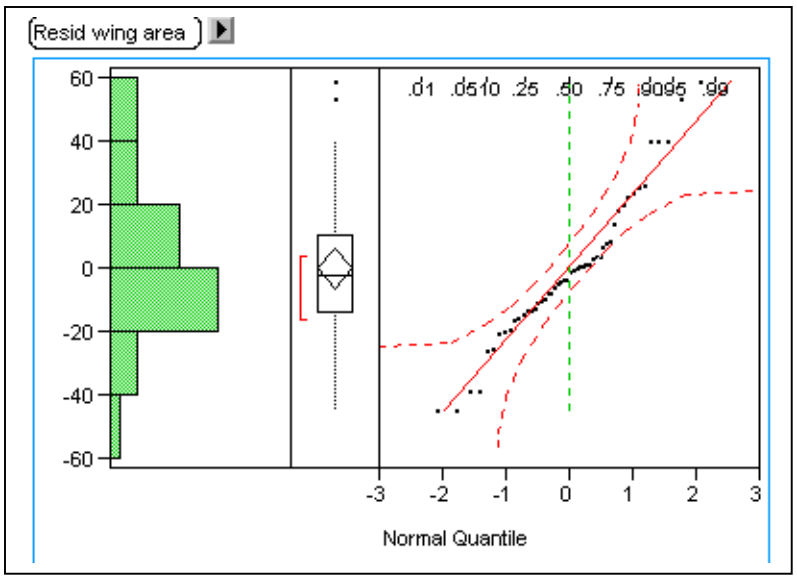
Histograms and Box Plots

Two common univariate techniques which could be used are the histogram and the box plot. Four scatter plots are presented below, together with a histogram and box plot of the residuals.

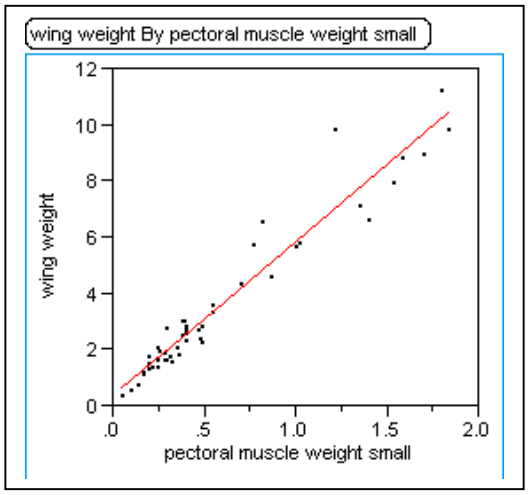


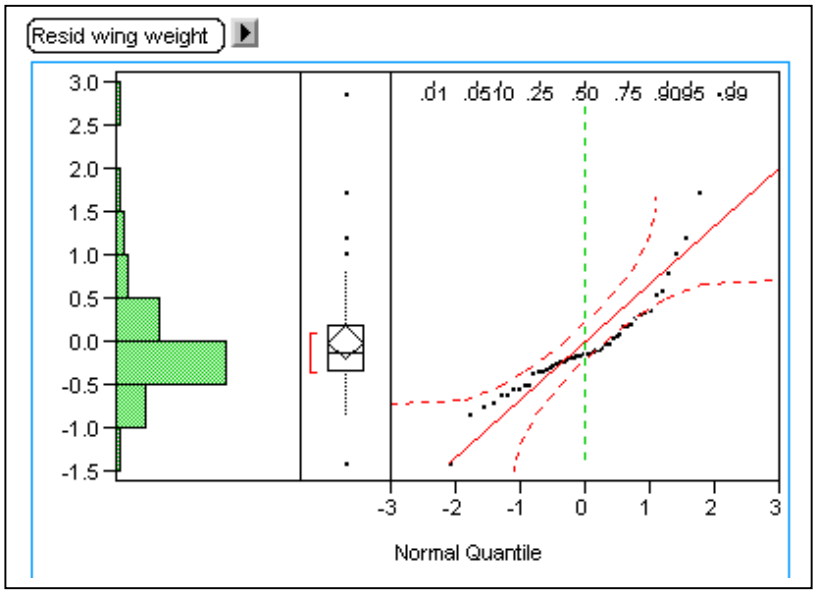


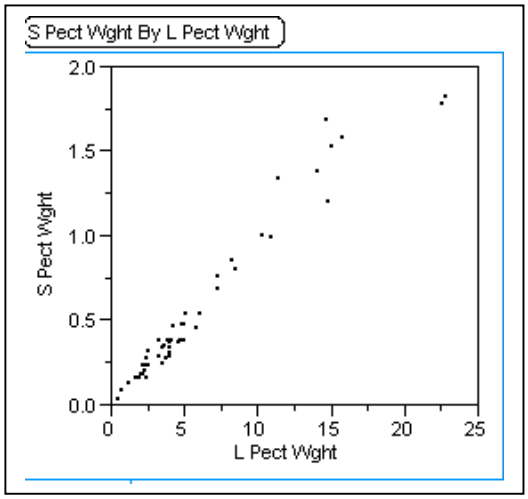


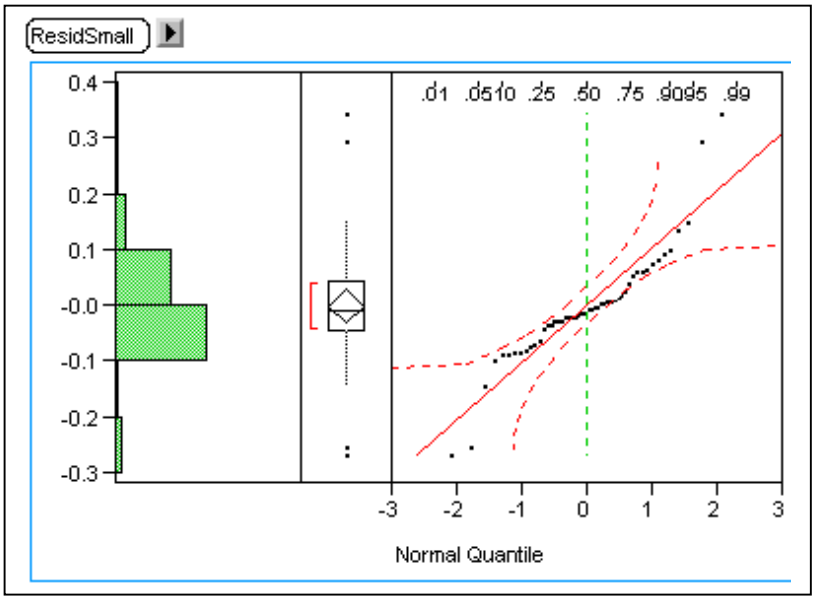


In the plots of residuals from Large pectoral vs. Wingspread and Wing area vs total weight both the histogram and box plot indicate some skewness in the residuals. The residuals are grouped toward the right rather than symmetric in appearance. The box plot shows two outliers in each case. Thus, we should be skeptical that an assumption of normally distributed residuals is warranted and investigate further.







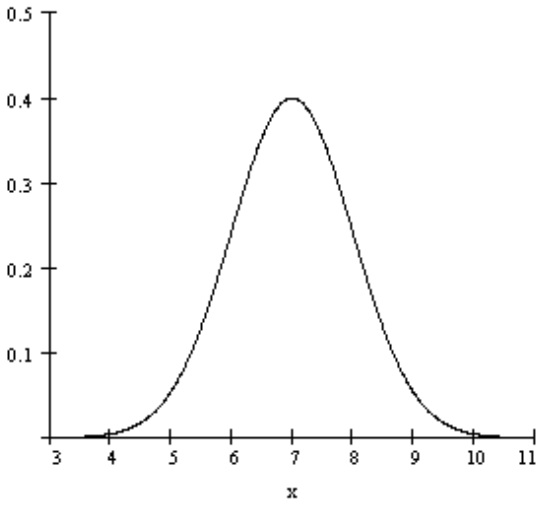


In the plot of residuals from wing weight vs. small pectoral muscle weight the residuals exhibit a more pronounced skew and the distribution is less variable than an approximately normal set of residuals. In the plot of residuals from Large pectoral vs small pectoral the residuals do not appear skewed, but are less variable than an approximately normal set of residuals. The pattern of outliers suggests the distribution

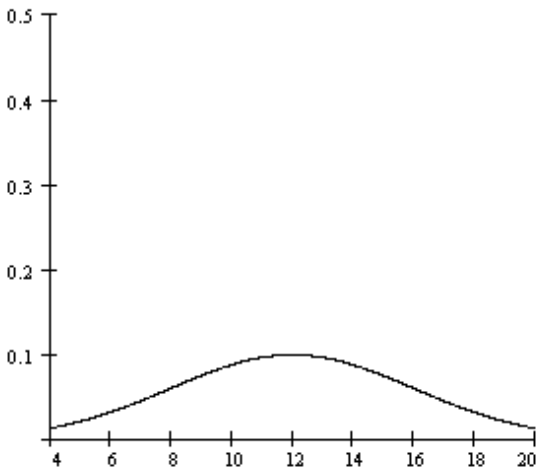
of residuals has fewer numbers in its tails than one might expect of approximately normally distributed data.

The normal probability plot

Another technique for checking the normality of data, available on most computer software and some calculators, is the normal probability plot. The normal probability plot is designed to assess the "match" between two distributions. Suppose, for example, we want to "match" two normal distributions. (We will assume from at the start they are normal.) Our strategy is engagingly simple; match a set of percentiles from the two different distributions. For the two normal distributions below, the first, second, and third quartiles have been chosen arbitrarily to be matched.



Normal Distribution A



Normal Distribution B

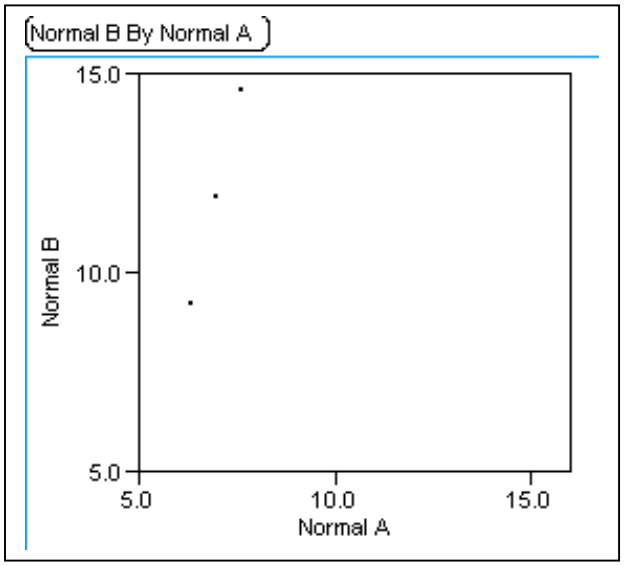
$\mu = 7$ $\sigma = 1$

$\mu = 12$ $\sigma = 4$

Using the normal curve tables we can find the correspondences between the quartiles and the z-scores; then we find the scores in the distributions which correspond to the Z-scores.

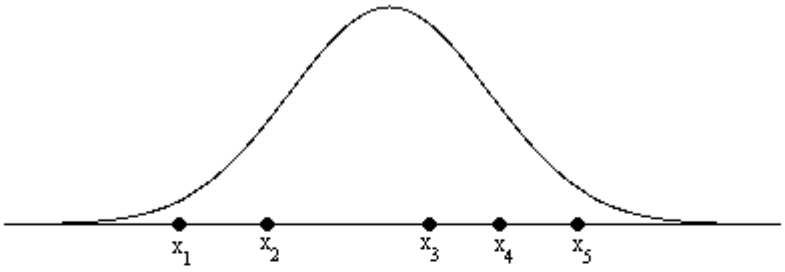
Quartile	Z-Score	Quartile Dist A	Quartile Dist B
1	-0.675	6.325	9.300
2	0.000	7.000	12.000
3	+0.675	7.675	14.700

Now construct a scatter plot of the points
(6.325, 9.300), (7.000, 12.000), (7.675, 14.700) . It is no accident that we see a straight line. In constructing this plot we are taking advantage of the following fact: if two distributions are the same shape (i.e. geometrically similar) a scatter plot of points of their corresponding percentiles will be a straight line.

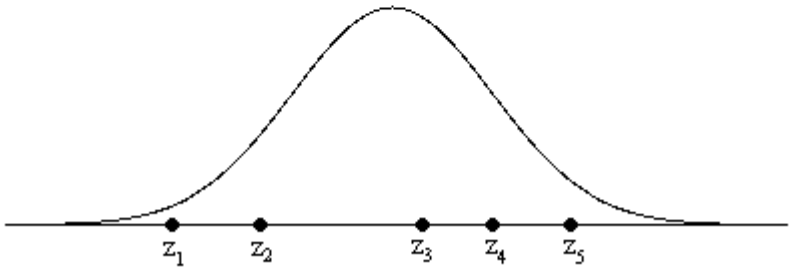


We could have put distribution B on the horizontal axis; had we done so the location of the points would change but they would still be on a straight line. The choice of which axis to assign to which distribution does not affect our interpretation of the plot in any way.

When matching the percentiles of actual data to the percentiles of a normal curve the standard normal curve is usually chosen to take advantage of the easy transformation from percentile to z-score afforded by the normal curve chart. We will use z-scores in the examples to follow. The motivation for the plot can be seen by appealing to the diagram below. (In actual practice we would have a much larger number of data elements.) For this example we have 5 data points, sorted from small to large. Notice also that 5 z-scores are positioned under the normal curve.



Normal Distribution with Data



Normal Distribution with z-scores

To construct the normal probability plot the points $(x_1, z_1), (x_2, z_2)$, ... (x_5, z_5) will be plotted and visually inspected to see if they fall on a straight line. But before we can do this we need to decide how to pair a z_i with an x_i. It turns out that pairing of data to a distribution is not as straightforward as pairing numbers from continuous distributions, for three reasons:

First, a discrete or finite set of data can never be exactly normal; at best it can be only approximately normal. Because of this, even data that is very close to being normally distributed will produce a normal probability plot, which is not exactly straight.

Second, data is a typically a sample from some population and the results of sampling can vary. Therefore a normal probability plot represents only an estimate of what the plot would look like were we to somehow "know" the actual distribution from which we are sampling. The existence of variability due to sampling should increase our caution when interpreting normal probability plots.

Third, statisticians are not completely in agreement about the best method of responding to the variability due to sampling. This means that statistics books, computer software, and calculator results can and will differ in how they assign a z_i to correspond with x_i. This should not cause undue concern for the reader – the differences among the various methods for matching the percentiles will produce very similar plots except in the case of small data sets. While it may be disconcerting to the student that her calculator, software, and textbook might produce three different answers for the same problem, she should realize these differences are insignificant when actually interpreting a normal probability plot. Each of the various methods will produce a plot of about the same straightness. Without further preamble, we will define the method for assigning percentiles to data elements and thus define the z-scores corresponding to a data set of n elements. (Devore & Peck, 1997)

Definition:

Order the n sample observations from smallest to largest. The i th smallest observation in the list is taken to be the

$$\left[\frac{100(i-0.5)}{n}\right]\textbf{Th sample percentile.}$$

Note: Other common definitions (Kimball, 1960):

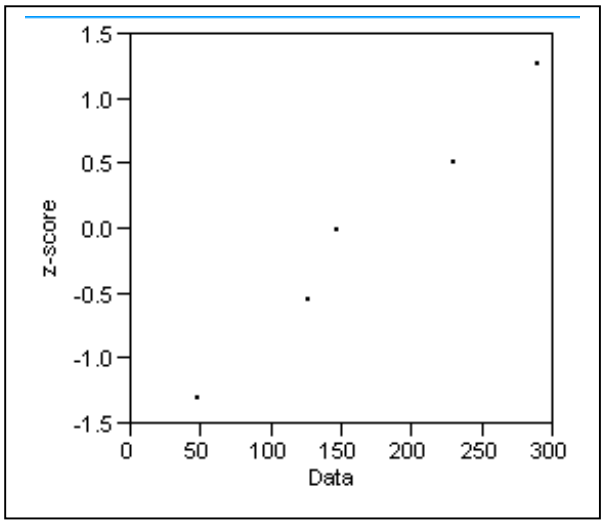
a) $\frac{100i}{n}$ b) $\frac{100i}{n+1}$ c) $100\left(\frac{i-\frac{3}{8}}{n+\frac{1}{4}}\right)$

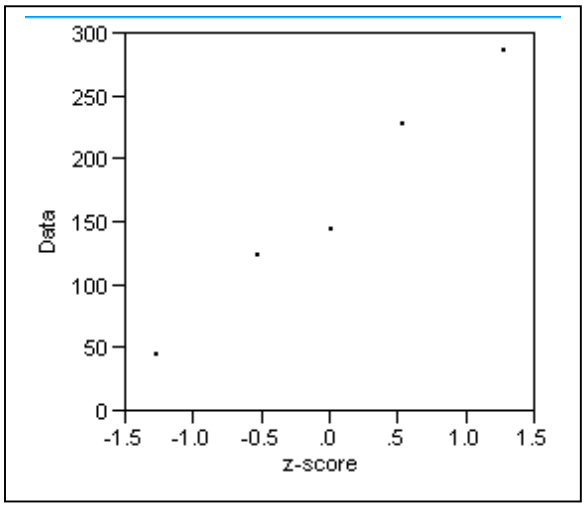
This algebraic definition can be understood graphically by appealing to the earlier diagram. We can think of the 5 data points as dividing up the normal distribution into 5 parts, each containing 20% of the area under the standard normal curve. $n-1$ of these parts, in this case 5 – 1 = 4 of them occupy the middle four fifths of the area; the fifth part is split up between the lower and upper tails of the normal distribution.. Thus, the areas labeled B, C, D, and E each contain 20% of the area under the curve, and the areas A and F together contain 20%.

Using our definition of the i th sample percentile above, we assign percentiles and use the normal curve table find z -scores that will be used for the normal probability plot. The table below summarizes the calculations.

i	Data element x_i	Data element (ordered) x_i	Sample percentile p_i	Corresponding z -score z_i
1	47.1	47.1	10	-1.28
2	126.0	126.0	30	-0.525
3	289.0	146.6	50	0.00
4	146.6	229.0	70	+0.525
5	229.0	289.0	90	+1.28

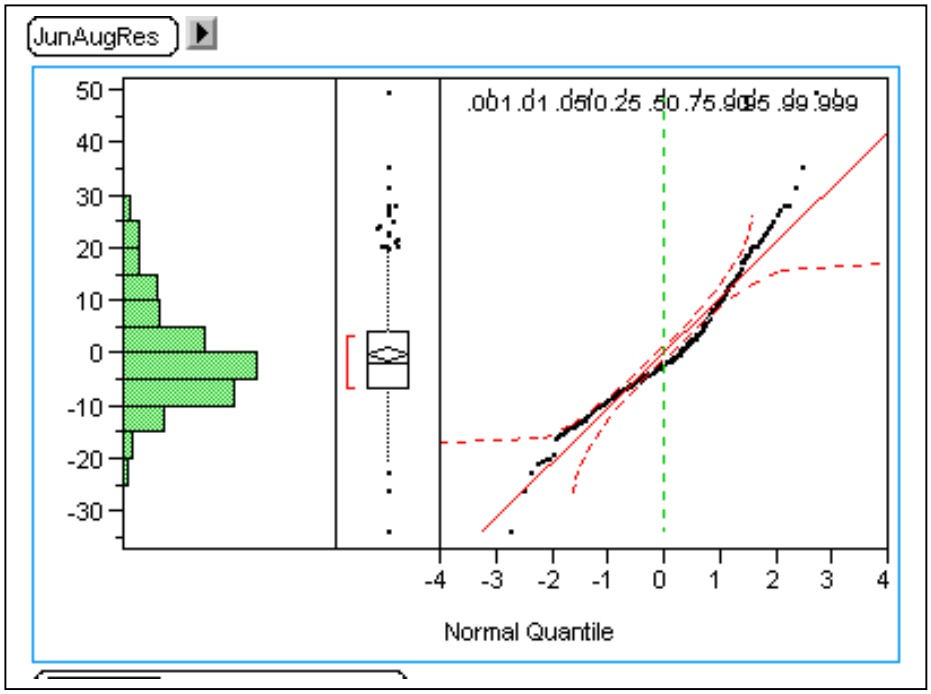
The normal probability plot is produced below. Notice that whether the original data or the z -score is placed on the horizontal axis is unimportant; the shape of the plot with the z -scores on the horizontal axis is just a mirror image of the plot with the original data on the horizontal axis.



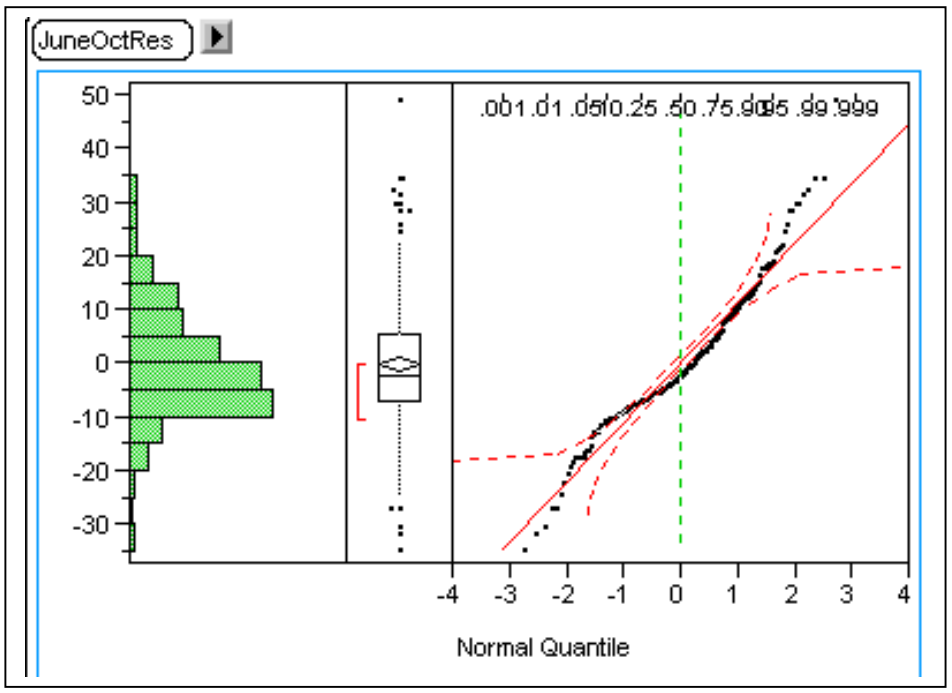


The advantage of the normal probability plot over a histogram and box plot is that the normal probability plot generates a more fine grained picture, whereas the histogram gives a more "course" and possibly erroneous view. The coarseness of the box plot is the result of using only five numbers to summarize the distribution. The histogram is more detailed, but a large cell width can disguise some relevant features of the data. The normal probability plot preserves every data point graphically and can provide a more detailed picture of the shape of the data.

The interpretation of a normal probability plot is a largely a matter of deciding whether or not the plotted points lie along a straight line. This is often a matter of judgement and experience, especially with small data sets. Often the inexperienced data analyst is unsure what he is looking for, but there are very few solid pieces of advice that work for all situations. It is generally true that unless a normal probability plot reveals obvious problems it is pretty safe to regard a data set as plausibly from a normal distribution.



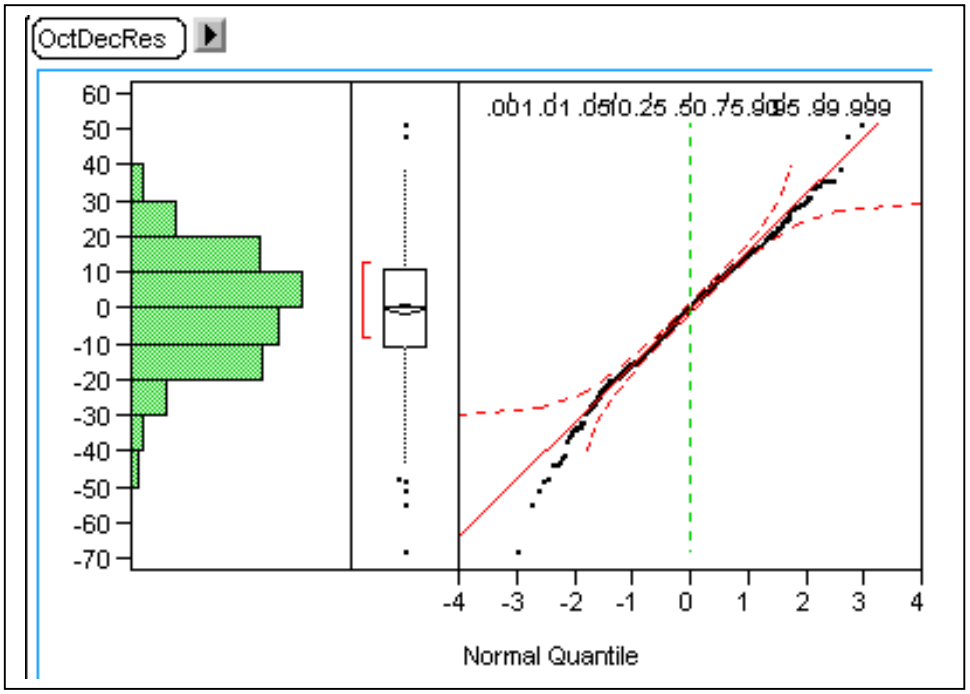
The plots at right, taken from the Confederate company sizes mentioned previously, provide examples of plots which clearly indicate potential problems as well as assist understanding of how a normal probability plot relates to the histogram and box plot. The original scatter plots are not shown; each plot represents about 500-600 data points.



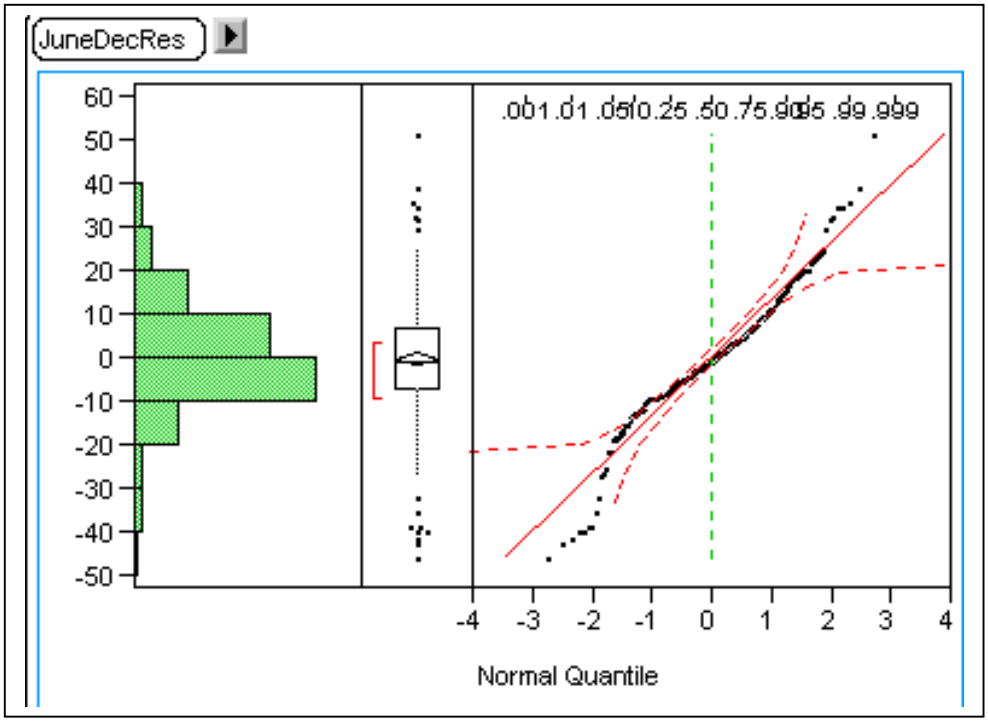
The first three sets of plots indicate data that, while not perfectly normal, seem close enough to use statistical

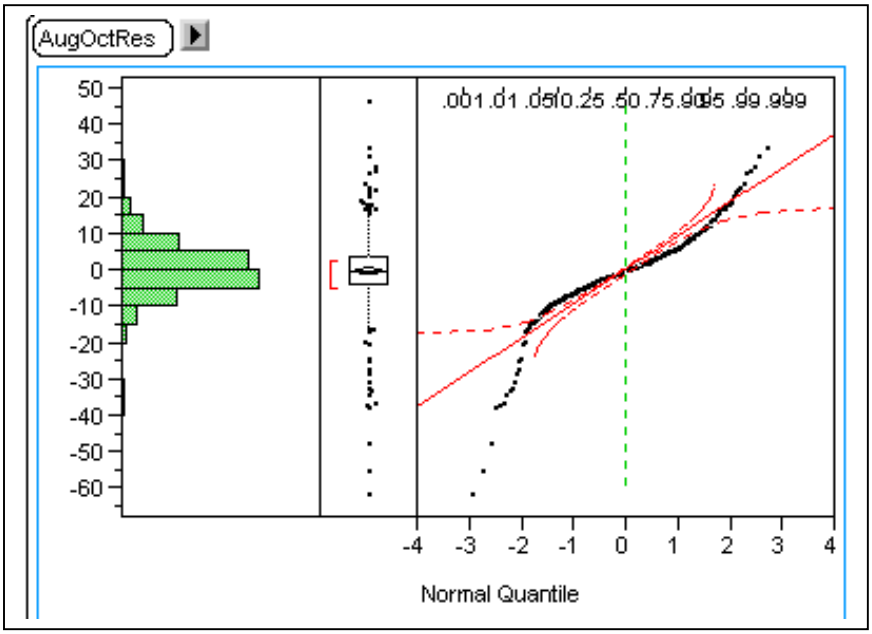
techniques that depend on normality.

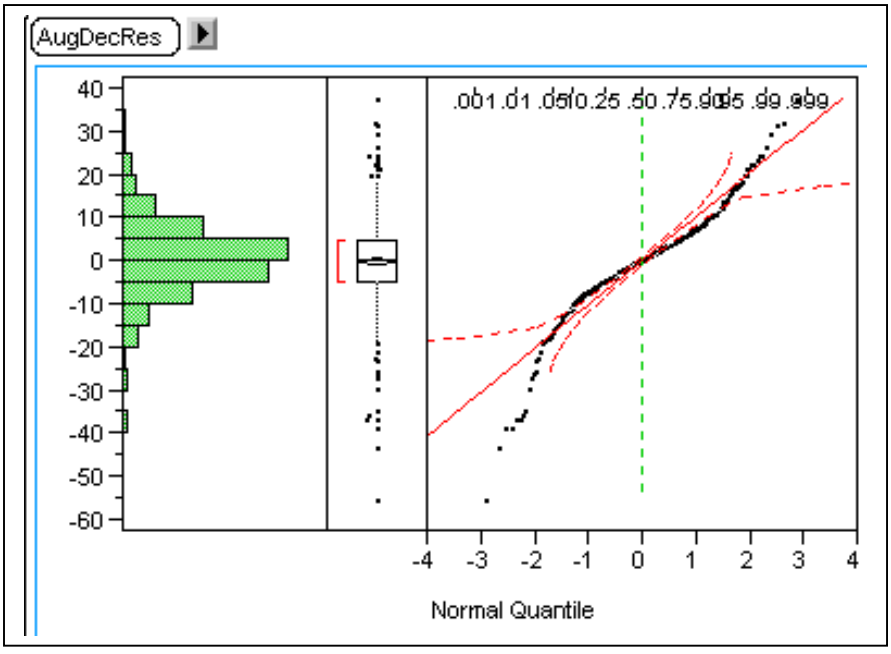
The histograms and boxplots indicate a slight skew, but generally the distributions are almost symmetric and the normal probability plots are nearly straight.



The next three sets of plots show residuals that are clearly problematic. In each case the deviation from a straight line by the normal probability plot is distinct and sudden. Notice also that neither the histogram or box plot makes this departure from normality as obvious as the normal probability plot.




AugOctRes
50
40
30
20
10
0
-10
-20
-30
-40
-50
-60
.001 .01 .05 .10 .25 .50 .75 .90 .95 .99 .999
-4 -3 -2 -1 0 1 2 3 4
Normal Quantile


AugDecRes
40
30
20
10
0
-10
-20
-30
-40
-50
-60
.001 .01 .05 .10 .25 .50 .75 .90 .95 .99 .999
-4 -3 -2 -1 0 1 2 3 4
Normal Quantile

Critical Values for Normal
Probability Plot Correlations

It turns out that the garden-variety correlation coefficient can be used as a tool to help assess the "straightness" of the normal probability plot for those borderline cases when the plot is not very clear, or at least one is unsure. The development and calculation of critical values for normal probability plots is described in a paper by Thomas A. Ryan, Jr., and Brian L. Joiner, located at the Minitab website, www.minitab.com. What follows summarizes part of the that paper, "Normal probability plots." (Note: this calculation should **NEVER** substitute for looking at the normal probability plot -- the plot can reveal patterns in the data that no single number could hope for!!) The idea behind the calculations is that if the correlation between the ordered data and the "expected *z*-score" is sufficiently less than 1.0, then the normal probability plot is sufficiently not straight to raise the suspicion of nonnormality in the population whence hence it came. "Sufficiently less than 1.0" is defined in a manner similar to the use of any inferential statistic, i.e. with an appeal to the question, "what is the probability this number would have arisen by chance alone even though the population is normal?" As with such statistics as the *t* or Chi-square, these numbers are translated into probabilities and assessments of credibility are made by appealing to critical values (in the case of a hypothesis test) or a p-value. Ryan and Joiner, using simulation methods, have provided methods for calculating the approximate critical values for *r*.

The approximate critical values of the correlation between normal scores and ordered data generated by Monte Carlo simulation were obtained for various *n*. Five hundred independent random samples for each value of *n* between 11 and 77, and 3500 samples for each *n* between 3 and 10 were generated. The empirical critical values were computed for $\alpha = 0.10,\ 0.05,$ and 0.01. These results were then smoothed for each of the three alphas, and a function of the following form fitted:

$$CV\left(n\right) = b_0 + \frac{b_1}{\sqrt{n}} + \frac{b_2}{n} + \frac{b_3}{n^2}\,.$$

The simple approximation formulae give critical values accurate to within the limits, ± 0.007 for α=0.10, ± 0.005 for α=0.05, and ± 0.002 for α=0.01.

These approximate critical values are given by the following formulae:

$$CV(n) = 1.0071 - \frac{0.1371}{\sqrt{n}} - \frac{0.3682}{n} + \frac{0.7780}{n^2}, \text{ for } \alpha=0.10$$

$$CV(n) = 1.0063 - \frac{0.1288}{\sqrt{n}} - \frac{0.6118}{n} + \frac{1.3505}{n^2}, \text{ for } \alpha=0.05$$

$$CV(n) = 0.9963 - \frac{0.0211}{\sqrt{n}} - \frac{1.4106}{n} + \frac{3.1791}{n^2}, \text{ for } \alpha=0.01$$

For n = 5, the formulae evaluate as follows:

$$CV(n) = 1.0071 - \frac{0.1371}{\sqrt{5}} - \frac{0.3682}{5} + \frac{0.7780}{5^2} = 0.903, \text{ for } \alpha=0.10$$

$$CV(n) = 1.0063 - \frac{0.1288}{\sqrt{5}} - \frac{0.6118}{5} + \frac{1.3505}{5^2} = 0.880, \text{ for } \alpha=0.05$$

$$CV(n) = 0.9963 - \frac{0.0211}{\sqrt{5}} - \frac{1.4106}{5} + \frac{3.1791}{5^2} = 0.832, \text{ for } \alpha=0.01$$

For n = 10, the formulae evaluate as follows:

$$CV(n) = 1.0071 - \frac{0.1371}{\sqrt{10}} - \frac{0.3682}{10} + \frac{0.7780}{10^2} = 0.935, \text{ for } \alpha=0.10$$

$$CV(n) = 1.0063 - \frac{0.1288}{\sqrt{10}} - \frac{0.6118}{10} + \frac{1.3505}{10^2} = 0.910, \text{ for } \alpha=0.05$$

$$CV(n) = 0.9963 - \frac{0.0211}{\sqrt{10}} - \frac{1.4106}{10} + \frac{3.1791}{10^2} = 0.880, \text{ for } \alpha=0.01$$

Ryan and Joiner, in their study, matched the *i*th smallest observations with the

$$p_i = 100\left(\frac{i - 3/8}{n + 1/4}\right) \text{th sample percentile.}$$

Using this definition of the *i*th sample percentile, we get the following results for the example above:

	The Data			The Theoretical	
i	Data element x_i	Data element (ordered) x_i		Sample percentile $p_i = 100\left(\frac{i - 3/8}{n + 1/4}\right)$	Corresponding z-score z_i
1	47.1	47.1		11.91	-1.1798
2	126.0	126.0		30.95	-0.4972
3	289.0	146.6		50.00	0.0000
4	146.6	229.0		69.05	0.4972
5	229.0	289.0		88.10	1.1798

The correlation between the ordered data elements and the expectations if the data were "normal" is: $r = 0.9909$. Since this is greater than any of our critical *r*'s above, we judge these data as having reasonably come from a normal population.

A TI-83 sequence for these calculations

The data: 289.0, 47.1, 126.0, 146.6, 229.0

1. **Construct a sequence of 1,…,n in List1**

   seq(x,x,1,n)→L1 (Generic)

   seq(x,x,1,5) →L1 (For *n*=5)

2. **Define Y1(X)=invNorm(x-0.375)/(n+.25))**

   Y1(X)=invNorm(x-0.375)/(5+.25)) (For *n*=5)

3. **Construct the normal scores in List2**

   Y1(L1) →L2

4. **Enter the data in List3**

5. **Sort (in place) List3 from small to large.**

6. **SortA(L3)**

7. **Make the scatterplot of List2 and List3 (ALWAYS LOOK AT THE GRAPH!!!)**

8. **Calculate *r*=0.9908988331.**

9. **Ask, "Is that *r* sufficiently out of line, to coin a phrase???"**

Detecting unequal variance and a non-linear pattern: the residual plot

Previously the normal probability plot was used as a check on the normality of the residuals in regression analysis. In addition to plausible normality of residuals a good linear fit exhibits a random pattern in the residuals when they are plotted against the explanatory variable. The graphic display commonly used to check for such a pattern in residuals is a scatter plot known as a "residual plot."

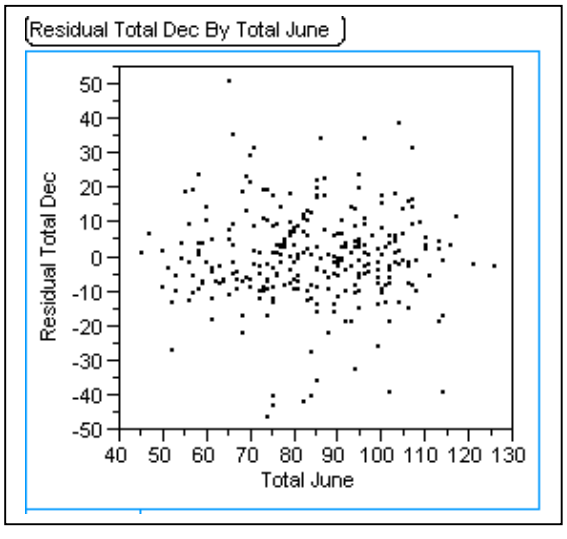
There are two varieties of residual plots:

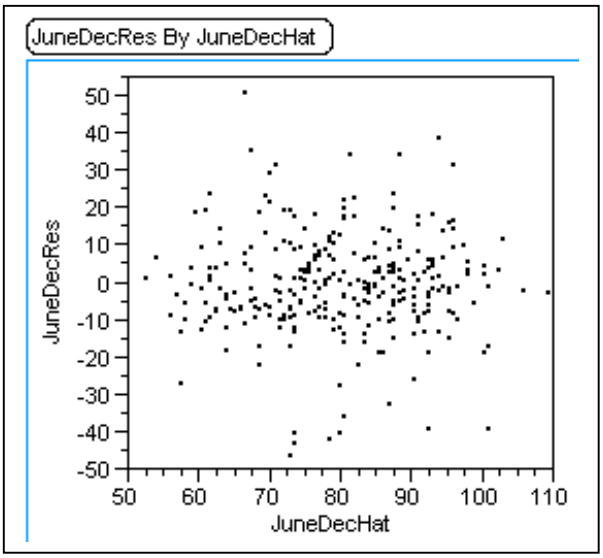
a) residuals are plotted vs. the explanatory variable

b) residuals are plotted vs. the predicted value $(\hat{y}_i)$ calculated using the best fit equation.

There is a difference between these two varieties of residual plots, but for purposes of simple regression diagnosis the difference is of no consequence. The "pattern" of residuals will be the same whether the explanatory variable or the predicted value is used on the horizontal axis. The interpretation of a residual plot proceeds by searching for a pattern in the plot which distinguishes it from what the plot should look like if the regression line is a good fit. To see what a good-fit lack-of-pattern looks like consider the scatter plot and best fit line for the relationship between Confederate company sizes in December and June, plotted earlier.

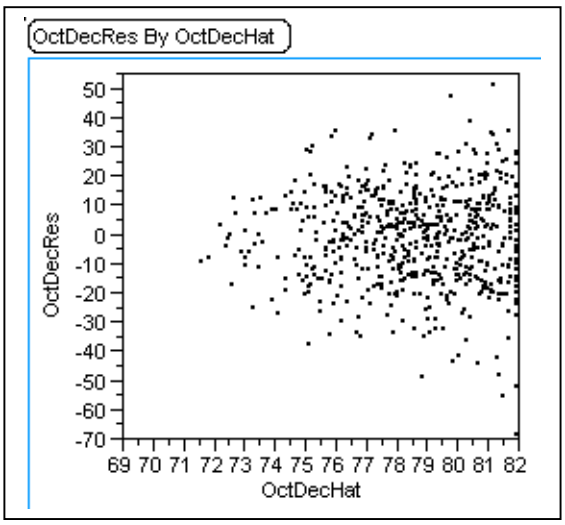
The residual plot with the explanatory variable (June company sizes) on the horizontal axis is plotted below on the left, and the residual plot with the predicted value (Predicted December company sizes) on the horizontal axis is plotted below on the right.

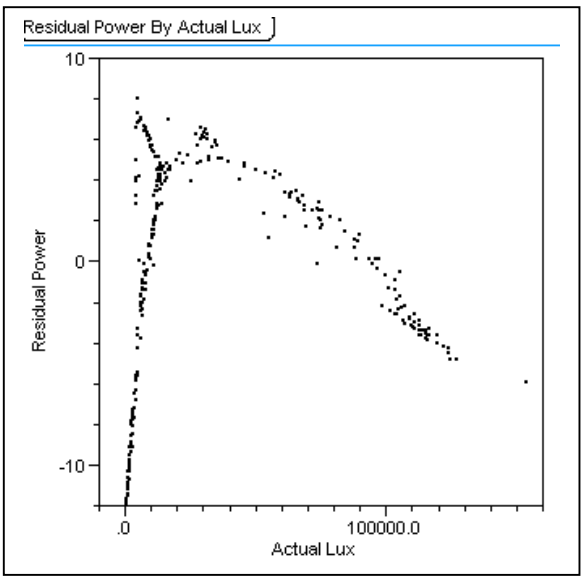




Notice that although the horizontal scales differ, the pattern of scatter is exactly the same. The points in these residual plots indicate there are more small residuals than large residuals, the sizes of the residuals is very similar all the way across the plots horizontally., and the pattern of points does not appear to curve in any way. The residual plot looks very much like the original scatter plot if the residuals are plotted against the explanatory variable. In fact, if we were to mentally twist the best fit line to a horizontal orientation and re-scale the vertical axis as deviations from the mean of the response variable we would have the residual plot!

The residual plot has psychological advantages over the original data plot with the best fit line overlaid. It is often easier to detect departures from fit by having a horizontal line rather than the best fit line as a reference. The standard parallelogram scatter of points for a good fit becomes a rectangular shaped scatter in the residual plot. We have previously, in Table X, listed the characteristics of both good and ill fits of a regression line. We will now present some symptoms of ill fit in the context of residual plots. The residual plots below all exhibit symptoms that the best fit line regression may not be the best technique for the respective data:

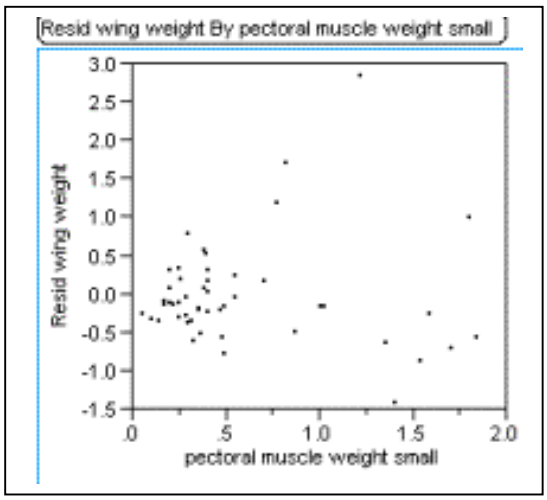


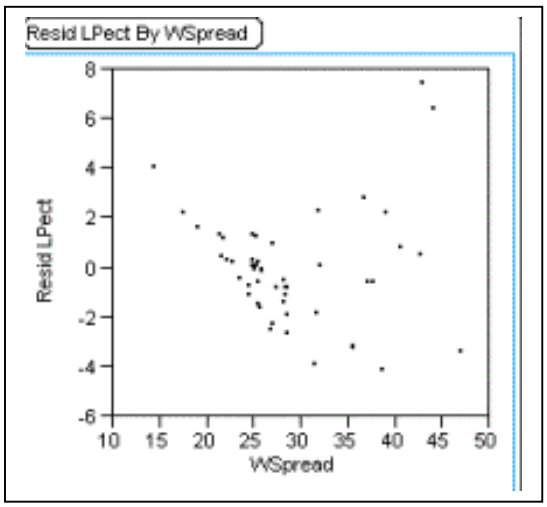


The residual plot from regressing the December company sizes on the October company sizes is a very good example of the dissimilar spread of residuals for different values of the explanatory variable. This characteristic is known as heterogeneity of variance, and is distinguished from homogeneity (sameness) of variance. When a plot exhibits homogeneity of variance, as do the plots of the residuals from the December vs. June company sizes above, the mind's eye can almost draw a rectangle around the points in a residual plot. In the December vs. October residual plot, the mind's eye could draw a triangle. The variability of the residuals is less for small companies than it is for large companies. This is a pattern that spells trouble for the standard least squares regression technique. The plot of the solar panel residuals is another indication of problems: rather

than a horizontal rectangular pattern to the residuals this plot exhibits a distinct curvature, indicating that the relation between the two variables might be more complex than initially thought.

The next two plots, Wing weight residuals vs. small pectoral, and large pectoral vs. wingspread, exhibit both curvature and heterogeneity of variance!





Bibliography

Chatterjee, S., & Price, B. Regression Analysis by Example (2nd). John Wiley & Sons, 1991.

Devore, J. & Peck, R. Statistics: The Exploration and Analysis of Data. John Wiley & Sons, New York. 1997.

Draper, N. & Smith, H. Applied Regression Analysis (3rd ed). John Wiley & Sons, New York. 1998.

Greenwalt, C. H. 1962. "Dimensional Relationships for Flying Animals." Smithsonian Miscellaneous Collections, v 144, No 2. April, 1962. From Table 15, Passereaux rameurs a vol soutenu.

Hamilton, L. C. Regression with Graphics: A Second Course in Applied Statistics. Duxbury Press, 1992.

Hamilton, C. H. "The trend of the marriage rate in rural North Carolina." Rural Sociology, post-1936.

Kimball, B. F. On the choice of plotting positions on probability paper. Journal of the American
Statistical Association. September, 1960.

Myers, R. H. Classical and Modern Regression with Applications (2nd ed). Duxbury Press, 1990.

Olsen, A., & Veenstra, J. Unpublished data.

Olsen, C. Unpublished data.

U.S. Dept of Interior, Geological Survey, Water Resources Branch, p 7-8. September, 1947.