Paper 274

Statistical Graphics

David A. Dickey
North Carolina State University
Raleigh, North Carolina

Abstract

The term "statistical graphics" can have several meanings. The most common use is a visual display of data possibly including some summary statistic such as a regression line running through a scatter plot. Another meaning, that which I hope to address herein, is a visual display that provides a clearer understanding of some statistical *principle*.

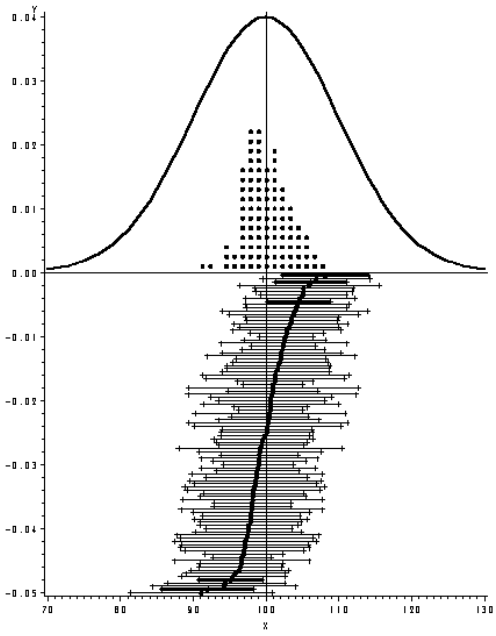
1. Introductory example

A nice example is Figure 1. This graph deals with confidence intervals. What exactly is a confidence interval? You read in your textbooks the formula for the calculation of a confidence interval. For example, after studying the properties of the normal distribution with mean μ you are told how to compute a confidence interval for the mean. You are told to compute the sample average $\overline{Y}$ then the confidence interval $\overline{Y} \pm t\sqrt{s^2/n}$. This gives two numbers L and U that serve as upper and lower bounds for the confidence interval.

Now that you know how to compute a confidence interval, what is it that you have computed, that is, what do the numbers mean? For example, if L=10 and U = 25 does that somehow mean there is a 95% probability that μ is between 10 and 25? Of course not - μ is a single (albeit unknown) number and does not vary at all so asking how often this single number is between 10 and 25 makes no sense. It either is or is not between 10 and 25. In fact, most well written textbooks make a point of saying that a confidence interval does not have a 95% probability of containing μ.

A well constructed graph should make the situation clear. In figure 1 you see a normal distribution displayed. Now suppose you select a sample of 10 observations from this interval, noting

Figure 1: Confidence Intervals



the mean and upper and lower limits U and L for this sample. You can draw a horizontal line connecting U to L just

below the normal curve with the sample average marked in the middle. Now you can select a second sample drawing its horizontal line below the first and continue selecting samples and drawing intervals in this fashion. The calculation of L and U is easily done in a data step and without using the annotate facility, the graph can be constructed. Let's add a couple of other informative features. First, for each sample, let us place a dot at the sample mean in the interval and inside the original normal curve, stacking dots whose coordinates round off to the same number so that the set of dots forms a histogram of the sample means. The histogram of dots should make these statistical points clear:

1. The sample means appear to have a normal distribution

2. The variance of this distribution is smaller than that of the parent population.

3. The mean of this distribution is the population mean.

4. The population mean is fixed - the sample means vary.

The second feature you might add is to sort your data set of Us and Ls in increasing order of sample means. In this way the horizontal line segments will illustrate some further points

1. The sample means have a distribution (the cumulative distribution function is the sorted sample means, their dots forming the characteristic S shape).

2. The widths of the intervals vary in a rather random fashion as you visually work your way up through the set of intervals. This illustrates the independence of the sample mean and variance (a result known as Basu's theorem).

3. The darker intervals fail to contain the population mean (which clearly does *not* vary). About 5% of the intervals miss.

4. The S shaped curve crosses the true mean reference line about half way up, so the sample means are median-unbiased (their median is the parameter being estimated)

This example graph shows clearly what is meant by 95% confidence as well as illustrating some other statistical concepts that can otherwise be a bit tough to grasp. In the sections that follow I will illustrate some further statistical ideas with graphics, this being the only unifying theme of the paper.

I will talk a bit about the programming for this first graph. The ideas are

1. Graph the population using the equation of the appropriate normal distribution. This is accomplished in a loop containing the equation.

2. Create the L and U endpoints for n intervals then, after

sorting by sample means, give interval i a Y coordinate like i/n -1 so that the intervals will stack below the horizontal axis for the normal curve.

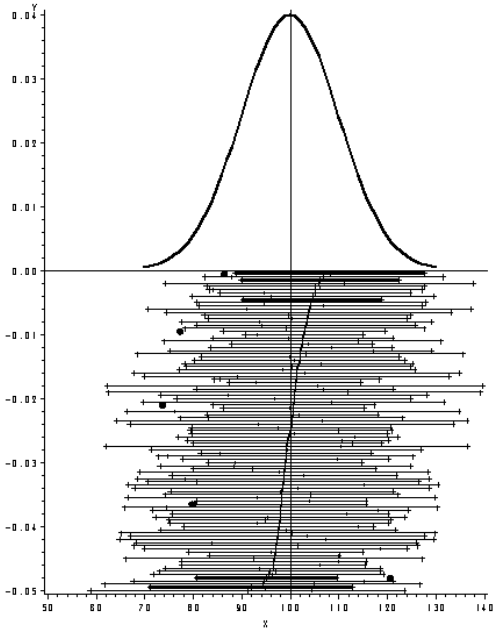
3. Use a different symbol to indicate intervals that do not contain the population mean. This is done by using an IF statement to check if an interval misses the population mean 100 then using an ID variable that is small for those that hit and large for those that miss the mean. Use the PLOT X*Y=ID syntax to pick plot symbols of different color. The entire graph is made using the ID variable to pick different plot symbols.

4. Make the histogram of dots inside the normal population. This is done by passing though the *sorted* list of sample means, rounding each to get the dot's X coordinate and setting the Y coordinate just above 0 if the X is not the same as its lag or else incrementing the Y coordinate to put the dot above its predecessor.

2. Prediction Intervals

While confidence intervals are a bit hard to understand, prediction intervals are even harder. Again, a good graph should help. For example, a 95% prediction interval is *not* an interval that will contain 95% of future values drawn from the distribution. Here is what actually happens 95% of the time. Draw a sample, compute a prediction interval, then select one more observation and see if it is in the interval. Now draw another sample, compute the interval, select one point and see if it is in the interval. Repeat this over and over. In 95% of the cases, the individual observation will be in the associated interval. The point and the interval change in each sample. To illustrate this idea, you use the same kind of idea as in section 1 but add the extra point using a different symbol when it misses the interval. This is Figure 2.

Figure 2: Prediction Intervals



Darker intervals appear for the samples whose confidence intervals miss the mean. Notice that these prediction intervals sometimes contain and sometimes miss the future point. The missed points are the large dots while small symbols mark additional points that fall into their predicting intervals. Note that the missed points are not necessarily in the 5% tails of the true population distribution.

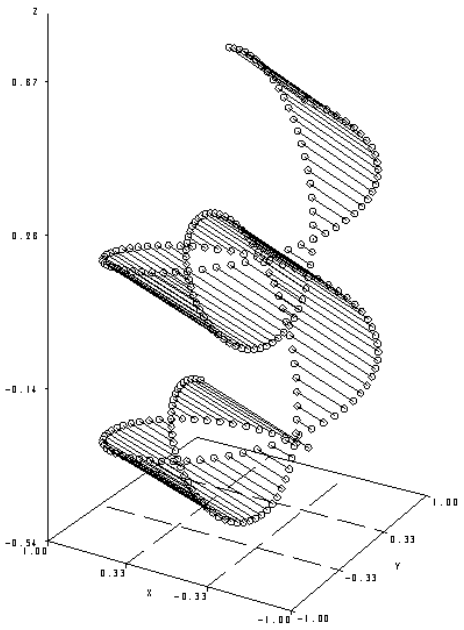
Finally notice that these intervals are of different widths. There is no way each of them can cover 95% of the population. Clearly that interpretation of a prediction interval (a given interval catching 95% of future values) is *wrong*.

3. Projections

In the theory of least squares regression you make heavy use of the matrix $\mathbf{X(X'X)^{-1}X' = P}$ which is a symmetric idempotent matrix(i.e. **PP**=**P**). *Any* symmetric idempotent matrix forms a projection onto the space spanned by its columns. Exactly what does this mean? To illustrate, you will use a data step to create a set of points in 3-dimensional space. These can be plotted in the SAS/GRAPH procedure PROC G3D or using SAS/INSIGHT. ™

The idea here is to produce a spiral of points that looks like a coil spring, then multiply the three dimensional vector of coordinates for each point by a symmetric idempotent matrix **P** to observe how this coil is projected into a lower dimensional space. In a data step, put a loop in which Z and angle A are incremented in each pass then let X=sin(A) and Y=cos(A) so that PLOT X*Y=Z; will produce a coil. Collecting the coordinates of each point in a column vector (X,Y,Z)' you can create a new point (x,y,z)' by multiplying your first vector by **P**. The rank of **P** will determine the dimension of the subspace into which you are projecting.

Figure 3



For a rank 2 matrix, the coil is projected into a 2-dimensional subspace and the result is dramatic if displayed in SAS/INSIGHT in which the original coil and projected coil can be rotated on the screen. The proper rotation of the projected coil will produce a line across the screen as you look at the edge of the plane into which you have projected. In Figure 3 I show the original coil, the projected version, and lines indicating how the projection moves point. The formula to get (x,y,z) from the original (X,Y,Z) is

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \frac{10}{14} & \frac{6}{14} & \frac{-2}{14} \\ \frac{6}{14} & \frac{5}{14} & \frac{3}{14} \\ \frac{-2}{14} & \frac{3}{14} & \frac{13}{14} \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

For example, if a point has coordinates (1,5,2) it becomes

™ SAS/INSIGHT and SAS/GRAPH are registered trademarks of SAS Institute, Cary, N.C.

$$\begin{pmatrix} \frac{36}{14} \\ \frac{37}{14} \\ \frac{39}{14} \end{pmatrix} = \begin{pmatrix} \frac{10}{14} & \frac{6}{14} & \frac{-2}{14} \\ \frac{6}{14} & \frac{5}{14} & \frac{3}{14} \\ \frac{-2}{14} & \frac{3}{14} & \frac{13}{14} \end{pmatrix} \begin{pmatrix} 1 \\ 5 \\ 2 \end{pmatrix}$$

Notice that if you regress **Y** on **X** with

$$X = \begin{pmatrix} 5 & 6 \\ 3 & 5 \\ -1 & 3 \end{pmatrix}, Y = \begin{pmatrix} 1 \\ 5 \\ 2 \end{pmatrix}$$

you get

$$b = (X'X)^{-1}(X'Y) = \begin{pmatrix} 35 & 42 \\ 42 & 70 \end{pmatrix}^{-1} \begin{pmatrix} 18 \\ 37 \end{pmatrix}$$

$$= \begin{pmatrix} -.4286 \\ .7857 \end{pmatrix} \qquad \hat{Y} = Xb = \begin{pmatrix} \frac{36}{14} \\ \frac{37}{14} \\ \frac{39}{14} \end{pmatrix}$$

You see that the projection of each point is the predicted value that you get by regressing on any multiples of the first two columns of the projections matrix (or any two of its columns for that matter). The columns of the X matrix give coordinates of two points. These points, (5,3,-1) and (6,5,3), along with the origin (0,0,0) are in the plane into which the points of the coil are projected , in fact these three points are enough to determine the plane completely.

In a similar way, not illustrated here, the coil can be rotated by multiplying each point by a 3x3 orthogonal matrix **R**. Recall that a matrix **R** is *orthogonal* if **R'R** = **I** (the identity matrix). This illustrates the mathematics underlying the ROTATE and TILT options in the SAS procedure PROC G3D.

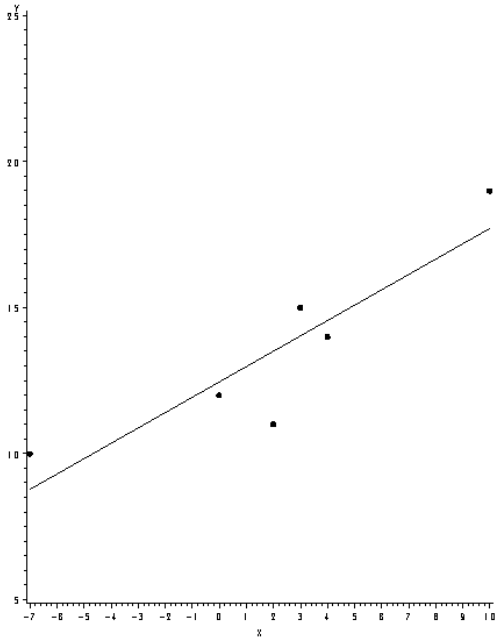
4. Confidence ellipses

Figure 4 is a scatter plot with 6 points. You know that when you run a regression line through this plot as

$$Y = \beta_0 + X\beta_1 + e$$

shown, you are fitting a model using the method of least squares, that is, you find the estimates b_0 and b_1 of the beta parameters that minimize the error mean square (MSE).

Figure 4: Regression



You also know that to test the null hypothesis that the betas are some specified values, such as

$$H_0 : \beta_0 = \beta_0^*, \beta_1 = \beta_1^*$$

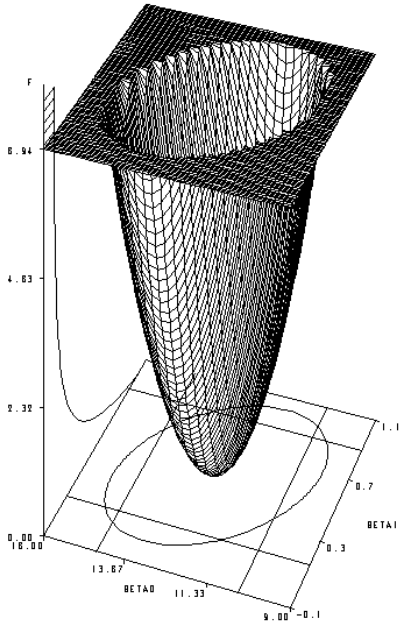
you will use an F test given by

$$F = \frac{(b_0 - \beta_0^*, b_1 - \beta_1^*) X'X (b_0 - \beta_0^*, b_1 - \beta_1^*)'}{2 * MSE}$$

and will reject the starred betas if F exceeds a critical value $F_{CRIT.}$

In figure 5, the floor is a grid of potential (β_0^*,β_1^*) values. Plot F as the height of each grid point, truncating to F_{CRIT} if $F > F_{CRIT}$. The result is a plane parallel to the floor with an elliptical hole in the middle. This is a 95% confidence ellipse for (β_0,β_1) if you take F_{CRIT} to be the 95th percentile of F.

Figure 5: F versus Betas



A parabolic "valley" descends beneath the ellipse, touching the plot floor at the least squares estimates (b_0, b_1). On the left wall of the plot is a picture of an F distribution with its upper tail (above the high plane) shaded. By mentally dropping the horizontal slicing plane down you can imagine the resulting decrease in ellipse area and associated decrease in confidence, measured by the area of the F distribution below the slicing plane.

On the plot floor, the 95% confidence ellipse is shown along with straight lines marking endpoints of the separate 95% confidence intervals for each parameter. The intersection of these forms a box, but clearly here is no reason to believe that the box has an associated 95% confidence.

Scheffe and Bonferroni, among others, have suggested methods for constructing simultaneous confidence limits for multiple parameters. The idea is to write down individual confidence intervals, one for each parameter, in such a way that all the confidence statements are simultaneously true in at least 95% of all samples.

To clarify, suppose you construct 20 confidence intervals, each a 95% confidence interval, from some experiment. For example, you might have a multiple regression with 20 parameters to estimate. Each interval misses its parameter in 5% of all samples. If the samples in which the first parameter is missed are completely different from those in which the second is missed then in 10% of all samples, at least one of these two parameters is missed. Thus with 20 parameters one could envision a scenario in which, for any run of the experiment, one of the intervals misses its target parameter.

Scheffe's idea is particularly relevant for the graph under discussion. He reasoned that intervals whose intervals intersect to form a rectangle that completely contains the 95% confidence ellipse would have at least 95% joint inclusion probability in repeated sampling. Thus he suggested expanding the confidence bounds until the ellipse is entirely enclosed by the rectangle.

For three parameters, the ellipse becomes a three dimensional ellipsoid

and the Scheffe method gives a box in three dimensional space. These topics are discussed in Rawlings et al (1998).

5. Cook's D

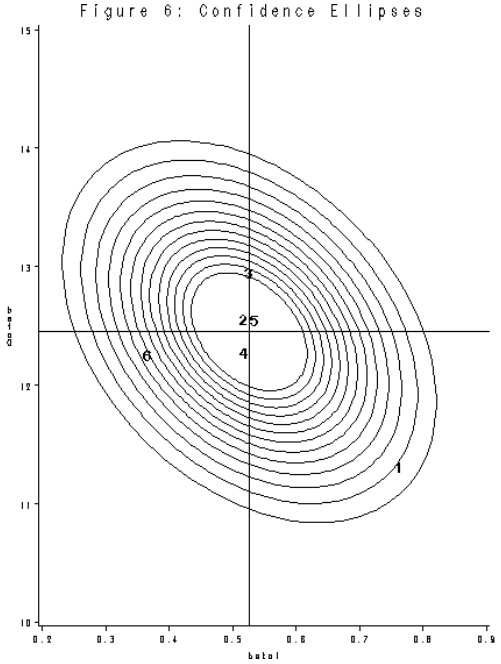
In figure 4, you saw a scatter plot and a regression line. Suppose you remove a point. What effect does that have on the regression line? Efforts to measure this effect fall under the general category of "regression diagnostics" and the specific subcategory "influence statistics."

One well known statistic, available in PROC REG, is Cook's D statistic. The formula for Cook's D and a general discussion of influence statistics and of other diagnostics can be found in Rawlings et al (1998).

When any individual point is removed from a regression, the parameter estimates will change from their original values. In the figure 4 data set they will change from the original least squares estimates (b_0 ,b_1) to something else, say (β_0 *, β_1 *) and you want a way of measuring the joint change in these two parameters.

To illustrate the idea of Cook's D, figure 6 shows the least squares estimates at the crosshair point, with concentric confidence ellipses surrounding it. The confidence ellipses shown have confidence levels from 20% to 80% in 5% steps but clearly if one plotted every conceivable confidence ellipse with confidence coefficients from 0 to 100%, these would fill the entire graph. In other words the new point (β_0 *, β_1 *) would have to lie on the edge of some confidence interval.

Figure 6: Confidence Ellipses



The formula for Cook's D is, from Rawlings et al (11.12)

$$D_i = \frac{(b_0 - \beta_0 *, b_1 - \beta_1 *)(X^{'} X)(b_0 - \beta_0 *, b_1 - \beta_1 *)}{2 * MSE}$$

which is seen to be just the formula for the F test for H_0:$\beta_0 = \beta_0^*$, $\beta_1 = \beta_1^*$ as given in section 4. The suggestion is to flag a point as influential if its removal produces a D bigger than some number so if you look up that number in an appropriate F table you see that the suggestion corresponds to flagging points as influential if they move beyond the edge of a certain confidence ellipse.

In figure 6, six (intercept, slope) pairs are plotted, corresponding to the removal of each of the six point in the plot of figure 4. Numbering the points in left to right order from 1 to 6 gives the plot symbol so you easily see that points 1 and 6 are the most influential and you have an idea, in terms of confidence intervals, the degree of influence for each.

A bit more detail is given by using PROC REG with the /R option on the MODEL statement to produce Cook's D. These can also be output to a data set with the OUTPUT statement. The problem here is that without recourse to an F table or inverse-F probability function, you would not have the probability interpretation. Since SAS has the appropriate probability function, it is easy to identify the confidence ellipse to whose edge you have moved. The following program does this for the small data set from figure 4. It uses PROC TIMEPLOT to display the results.

```
DATA SMALL;
INPUT X Y @@;
CARDS;
-7 10  0 12
 2 11  3 15
 4 14  10 19
     ;
PROC REG;
  MODEL Y=X / R;
  OUTPUT OUT=OUT1
      COOKD=COOKD;
DATA NEXT; SET OUT1;
PCT =
 100*PROBF(COOKD,2,4);
PROC TIMEPLOT;
   PLOT PCT;
   ID X Y COOKD;
RUN;
```

The PROC TIMEPLOT graph clearly points out the influential nature of points 1 and 6. The use of the ID statement lets us identify the points in question. In large data sets, it can be helpful to sort the data in the order of PCT so that the most influential points appear together, or to use a WHERE statement to see only those points whose removal shifts the parameters beyond a 25% ellipse.

Here is the PROC TIMEPLOT output, slightly reformatted to fit this document:

```
X  Y  D    PCT     min                max
                   *--------------------*
-7 10 2.05 75.60   |                    P|
 0 12 0.01  1.08   |P                    |
 2 11 0.28 22.74   |      P              |
 3 15 0.04  4.27   | P                   |
 4 14 0.02  1.63   |P                    |
10 19 1.02 56.22   |               P     |
                   *--------------------*
```

It is now seen that the removal of point 1 produces a D statistic 2.05 which is about at the 75th percentile of an F with 2 numerator and 4 denominator degrees of freedom. Thus removal of the first point would shift the parameter estimates out about to the border of a 75% confidence ellipse.

References:

Rawlings, J. O, S. G. Pantula and D. A. Dickey , Applied Regression Analysis, a Research Tool, Springer, New York, 1998.

Contact Information:

David A. Dickey,
Dept. of Statistics,
Box 8203,
N. C. State University,
Raleigh, N. Carolina 27695-8203.

dickey@stat.ncsu.edu