

USING GRAPHICS TO CONVEY STATISTICAL CONCEPTS AND RESULTS

D. A. Dickey, Department of Statistics, North Carolina State University

ABSTRACT

Many statistical ideas are nicely illustrated using SAS/GRAPH®. Results of experimental data can also be nicely conveyed using graphics. I will present an assortment of graphs relating to statistical analysis.

INTRODUCTION

There are many concepts in statistics that are nicely illustrated using graphics. In this paper, six fairly sophisticated statistical ideas are examined using SAS Institute's SAS/GRAPH software. All plots make heavy use of the ANNOTATE facility. These graphics are to appear in the third edition of Principles and Procedures of Statistics (Steel, et al).

DISTRIBUTION OF THE STANDARD DEVIATION

Most introductory statistics texts discuss the distribution of the sample variance, pointing out that it is an unbiased estimate of the population variance and that, upon normalization, it follows a Chi-square distribution. The sample standard deviation S , however, involves a nonlinear transformation of the sample variance. Is it still unbiased? What does its distribution look like?

To answer these questions, we draw 500 random samples, each of size 10, from a normal distribution with mean 40 and variance 144. For each sample, the sample standard deviation S is calculated and the resulting histogram is displayed in Figure 1. Note that the histogram rectangles are hung from the theoretical distribution of S rather than sitting on the horizontal axis. This allows the deviations from theory to be visualized as deviations from the straight horizontal axis rather than the curve.

How is the theoretical density plotted? We create a grid of 800 S values. For each S , its square properly

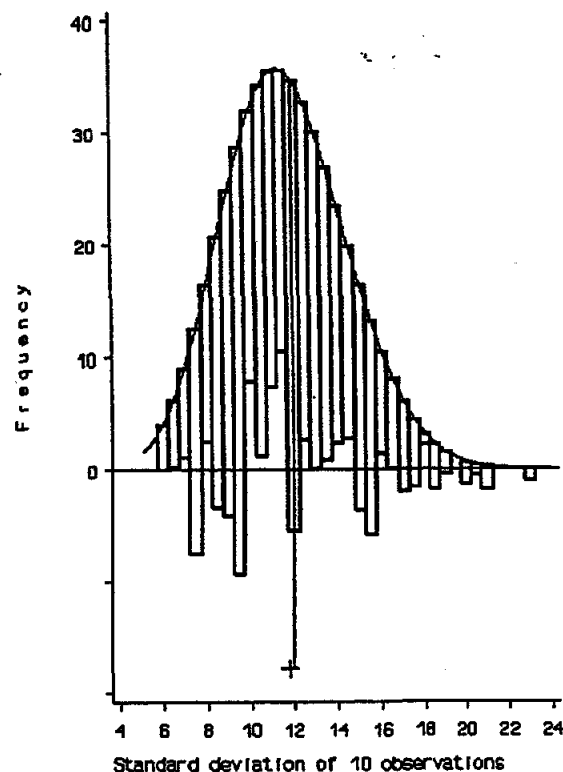


Figure 1. 500 sample standard deviations

normalized follows a Chi-Square distribution and so the probability of a value of S lying between two adjacent grid points is the same as the probability of its square lying between the squares of the grid points. We compute this probability using SAS® software's Chi-Square probability function and this gives the height of the theoretical curve and is then normalized to fit over the histogram based on our $n=500$ observed S values. A + sign appears at the mean of the 500 sample S 's and a vertical line appears at 12, the theoretical standard deviation. The average S is seen to be less than 12, illustrating the downward bias in S . The square root has produced a distribution much less skewed than the Chi-Square and in fact appearing somewhat normal.

PREDICTION INTERVAL FOR AN INDIVIDUAL

In statistics we distinguish between a confidence interval for the mean and a prediction interval for an individual observation. This can be in the context of a sample from one normal population or a regression. Although 95% confidence and 95% probability are related concepts, they are not the same. In Figure 2, we see a normal population with the middle 95% delimited. From this population, 100 random samples of size 10 were drawn and for each sample, a 95% individual prediction interval calculated. The intervals are then sorted by sample means (the interval midpoints) and are graphed below the population. The sample means trace out an S shaped and a moment's reflection convinces one that this is the empirical cumulative distribution function of the sample means. Comparing this to the normal population above shows the reduction in variance due to averaging - none of our 100 sample means are near the outer tails of the original population.

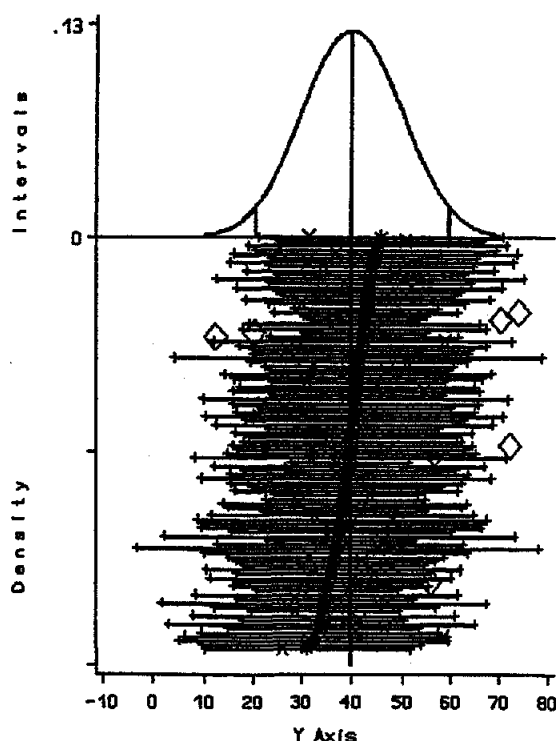


Figure 2. Prediction Intervals for Individuals
From Normal: Mean = 40, Sigma = 10

The interval endpoints form jagged edges for the ensemble of confidence intervals. This shows the effect of variation in S from sample to sample. Because the sample mean and variance are independent, the narrow and wide intervals seem randomly dispersed through

the ensemble of intervals. Clearly the mean and the confidence interval endpoints are random variables. For each sample, an eleventh point from the population is selected. This is plotted as an X if it falls in the interval and a large diamond if its outside. Roughly 5% of the values fall outside their respective intervals as is to be expected. Note that a given interval does not enclose 95% of the population and hence does not capture 95% of all future values from the population. Studying this graph helps us understand exactly what it is that happens 95% of the time.

CONFIDENCE INTERVALS AND REGIONS

A centered regression is run on two explanatory variables. We can find a 95% confidence interval for each slope. However, the probability that a sample will yield two confidence intervals each containing the appropriate slope parameter is not 0.95 nor is it 0.95^{**2} unless the columns of the X matrix are orthogonal. A 95% simultaneous confidence region for both slope parameters forms an ellipse.

To illustrate what is going on, let the model slopes be B1 and B2 and the estimates b1 and b2. Values of b1 and b2 are given by a regression program and the 95% confidence region is the set of all potential (B1, B2) points that cannot be rejected by the data, that is, it consists of all potential parameters (B1, B2) such that

$$(b1-B1,b2-B2)^T X^T X (b1-B1,b2-B2) / (2 \text{ MSE}) < F$$

where T stands for the matrix or vector transpose and F denotes the 95% critical point of an F distribution, MSE is the regression error mean square, and X is the nx2 centered design matrix of the regression.

To illustrate the idea, we create a 3 dimensional plot (see Figure 3) with B1 and B2 taking on values over a grid in the plot floor and with the height of each point being

$$\min\{(b1-B1,b2-B2)^T X^T X (b1-B1,b2-B2) / (2 \text{ MSE}) , F\}$$

so that the surface is truncated at the 95% F critical value and gives at the top, a plane with the confidence ellipse inscribed. The fact that this ellipse is almost but not exactly circular results from the near linear independence of the columns of X. Using ANNOTATE we project it into the floor as well. Straight lines in the floor denote the individual 95% confidence intervals. The F distribution is shown along the vertical axis with the upper 5% tail shaded. Recall

that the numerator of F is a Chi-square with only 2 degrees of freedom and hence follows an exponential distribution. That is why this F density function does not begin at the point (0,0) as do F 's with higher degrees of freedom. One can imagine sliding the top horizontal plane downward as the confidence level decreases and watching the corresponding shrinking of the confidence ellipse.

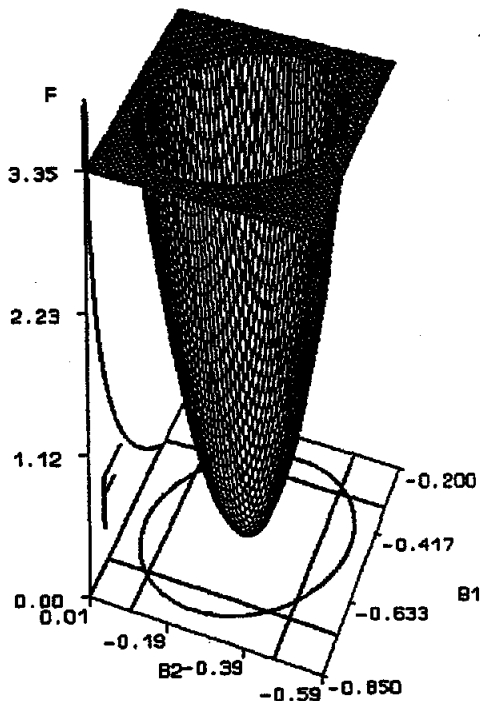


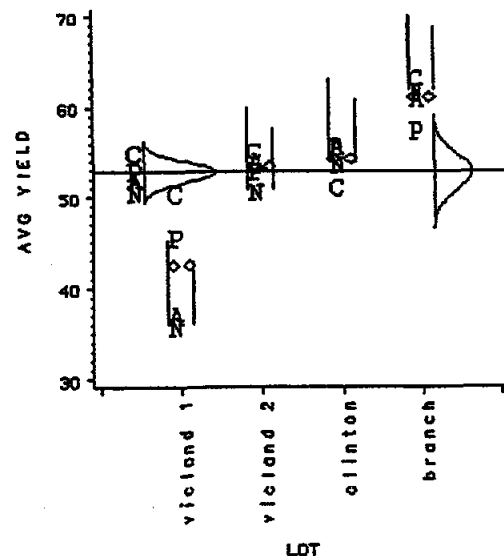
Figure 3. Confidence regions for (B1,B2)

A SPLIT PLOT ANALYSIS

In Principles and Procedures of Statistics, chapter 16, a split plot experiment is described in which 4 large plots of land called blocks are each divided into 4 pieces called whole plots. To each whole plot within a block, one of four varieties of oats is sown so that each block contains all four varieties: Vicland1, Vicland2, Clinton and Branch. Each whole plot is further subdivided into 4 split plots and one of four fungicides is applied to each of the split plots within a whole plot. The fungicides used are Ceresan M, Panogen, Agrox and None (i.e. an untreated check).

Because of the way the treatments are laid out, the experiment is a split plot and thus its analysis contains two different error terms, one for comparing whole plot means and one for comparing split plot means. In Figure 4 we see the 4 varieties listed on the horizontal axis and above each we have 4 large letters indicating the mean yields of oats for the 4 fungicides with that

variety. Because we have a control and several treatments, Dunnett's test is appropriate and lines from the letter N (untreated check) extend upwards an amount equal to the Dunnett 5% and 1% critical ranges. We extend only upward because we assume fungicides will not decrease yield. For each oat variety, a pair of diamonds denotes the variety mean yield. On the right side of the plot, a normal distribution extends three standard errors above and below the overall mean yield. This standard error is taken as the square root of $E_w/16$ where E_w is the whole plot error mean square and there are 16 observations averaged in each variety mean. Thus the spread in the diamonds is to be compared to this curve. On the left side, a similar normal distribution is shown that is appropriate for comparing fungicide means. These means are shown next to the curve. We conclude that Branch oats gives the highest yield and for Branch, no fungicide shows significant yield improvement.



A=AGROX, C=CERESAN M, P=PANOGEN, N=NONE (CONTROL)

Figure 4. Split plot analysis of oats data (Steel & Torrie)

EFFECTS OF LOGARITHMIC TRANSFORMATION

Statisticians often advise the use of the logarithmic transformation when data appear to have variation that increases as the mean increases, and when the distribution of the data appears positively skewed. It is implied that, on the logarithmic scale, the data will look more normal and homogeneous. To illustrate the effects of the logarithmic transformation, we present Figure 5. Here the curve $Y=\exp(X)$ is plotted using diamonds on a grid of equally spaced X 's. Thus $X=\ln(Y)$. Note that as the X 's get larger, the diamonds (and hence their Y coordinates) become

increasingly farther apart. Three normal curves labelled at their means are shown. They have the same variance but different means. On the rightmost normal curve there are two narrow strips of equal area, or probability, shown. Clearly $\Pr\{a < X < b\} = \Pr\{\exp(a) < Y < \exp(b)\}$ so that the area must be preserved on the exponential scale. Since the rightmost exponentiated interval end points are spread more than those of the left interval, the height of the Y density must be less for the right interval to preserve equal areas. The resulting curves are shown on the vertical axis and are labelled at their means.

Now let's reverse our thinking and imagine Y as original data. The distribution of Y is seen to be skewed with mean far to the right of the median. Viewing our three distributions on the Y scale shows obvious differences in variance which are undone by the logarithmic transformation that takes us from Y to X. Skewness is also undone by the transformation. It is fairly well known that if M is the mean of an X distribution, the mean of the Y distribution will exceed $\exp(M)$ and some practitioners try to apply a bias correction to the exponentiated sample mean of X. However, it is seen from the picture that the mean of the Y distribution may not always be the thing one wants to estimate since the labelled curve means have less than 50% probability to their right.

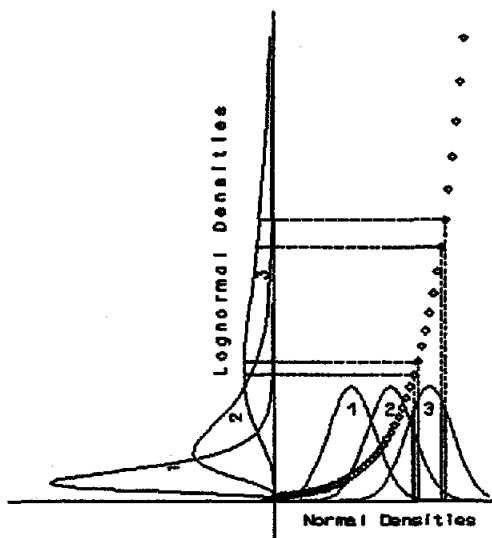


Figure 5. Effects of logarithmic transformation

CATEGORICAL DATA MODELING

One approach to the analysis of categorical data, especially binary data consisting of 0's and 1's, applies

analysis of variance or regression to some transformation of the data such as the logistic transformation. Using p to denote the probability of getting a 1, we may want to model p as a function of some variable X. For example, the probability p of a wound being healed may depend on time X. The idea is to model the logit $L = \log(p/(1-p))$ as $L = A + BX + e$ where A and B are the parameters to be estimated. In Figure 6, L is plotted on the vertical axis and p on a short axis parallel to the X axis. The S shaped curve shows how the logistic transformation maps the bounded variable p onto the entire L axis. A regression line $L = A + BX$ is shown on the graph. For any X, go up to the regression line so you are now at height L then move horizontally to the logistic curve then straight down to read the p associated with X.

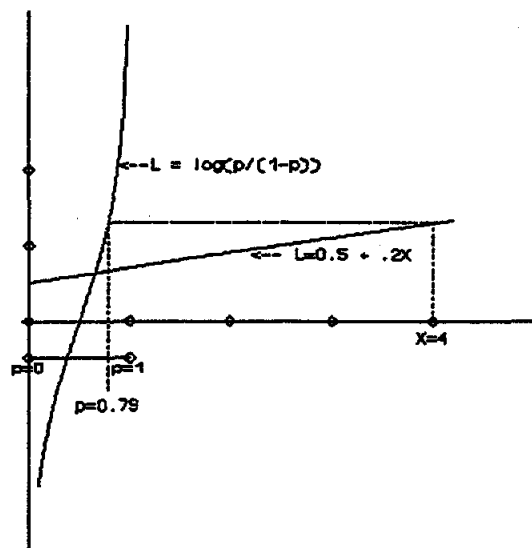


Figure 6. Logistic regression example

• SAS and SAS/GRAPH are registered trademarks or trademarks of SAS Institute, Inc. in the USA and other countries. • indicates USA registration.

REFERENCES

Steel, R. G. D., J. H. Torrie, and D. A. Dickey. Principles and Procedures of Statistics, 3rd ed. (in review) McGraw-Hill, San Francisco.