Letters to the Editor

Letters to the Editor publishes discussions of papers that have appeared in *The American Statistician* and of important issues facing the statistical community. Letters discussing papers in *The American Statistician* must be received within two months of publication of the paper; the author of the paper will then be given an opportunity to reply, and the letters and reply will be published together. All Letters to the Editor will be refereed. Corrections of errors noted in articles previously published in *The American Statistician* will be listed in Corrigenda at the end of this section.

SOME REMARKS ON QUICK ESTIMATION OF THE CORRELATION COEFFICIENT

Châtillon (1984) presents a simple and effective procedure for estimating the sample correlation coefficient (r) from a scatter diagram, based on vertical size measures of a "balloon" or ellipse sketched around the bulk of the plotted points. Theoretical justification is given for the case of samples from the bivariate normal distribution as well as for densities that are uniform on the interior of some ellipse. The following observations may be pertinent.

1. It is easy to see, by considering mixtures of the elliptical uniform densities mentioned above, that any random vector (X, Y) for which all level curves of its joint density are ellipses of the form $x^2 - 2\rho xy + y^2 = c^2$, where x and y are the values of the standardized (mean 0, variance 1) variables, has correlation $(X, Y) = \rho$. Thus whenever the scatter diagram has an elliptical appearance throughout, the balloon method should work well.

2. For the case in which the graph is scale standardized so that one standard deviation (sd) in the vertical direction encompasses the same distance as one sd in the horizontal direction, an alternative procedure for estimating r can be used: Sketch the ellipse as before. Let D and d represent the lengths of the principal axes of positive and negative slope, respectively. Then use

$$\hat{r} = (D^2 - d^2)/(D^2 + d^2). \tag{1}$$

As with Châtillon's formula, (1) is a sample version of the corresponding theoretical result that holds for bivariate densities with elliptical contours. The formula (1) has several nice properties: (a) The relevant quantity D/d is easy to estimate visually—it is not even necessary to draw the principal axes. (b) It is not necessary to compute a square root. (c) Many of the simple ratios $D:d$ result in easy decimal expressions for r (e.g., 2:1, 3:1, 7:1, 4:3).

A final remark on estimating correlation with ellipses: To check that a hand-drawn curve is indeed an ellipse, note that when viewed at an appropriate angle, a perfect circle should be seen.

Mark F. Schilling
Mathematics Department
University of California
University Park MC-1113
Los Angeles, CA 90089-1113

Reference

CHÂTILLON, G. (1984), "The Balloon Rules for a Rough Estimate of the Correlation Coefficient," *The American Statistician*, 38, 58–60.

REPLY

Schilling's first observation helps to understand why the uniform model (inside an ellipse) and the bivariate normal give the same formula to estimate the correlation coefficient, if the bivariate normal is considered to be the mixture of a great number of uniforms. In fact, we can demonstrate the following theorem, which is more general.

Theorem: Let (X, Y) be a couple of continuous random variables following the model

$f_1(x, y)$ inside a certain region R_1,
0 elsewhere.

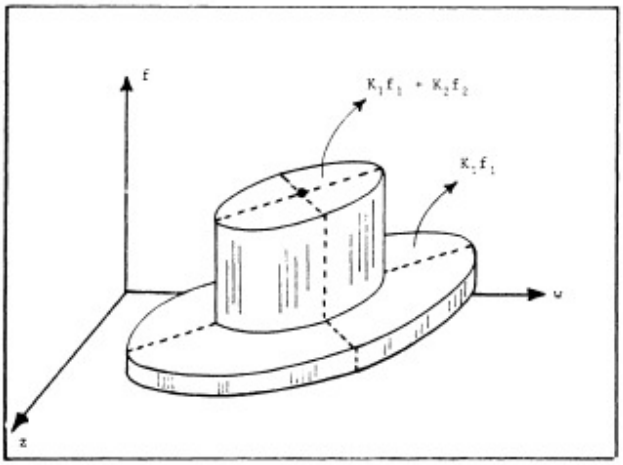

Figure 1. A Bivariate Density With a Hat Shape. Obtained by combining two uniforms with the same center and the same orientation. The correlation is preserved.

Let (U, V) also be continuous random variables following

$f_2(u, v)$ inside R_2,
0 elsewhere,

such that

$$E(X) = E(U)$$

$$E(Y) = E(V)$$

$$r(X, Y) = r(U, V) = r$$

$$\sigma_U = K\sigma_X,\ K > 0$$

$$\sigma_V = H\sigma_Y,\ H > 0$$

Let us combine those two distributions in the following way: (Z, W) is a random vector following

$$\begin{aligned} f(z, w) &= K_1 f_1 && \text{if } (z, w) \in (R_1 - R_2) \\ &= K_2 f_2 && \text{if } (z, w) \in (R_2 - R_1) \\ &= K_1 f_1 + K_2 f_2 && \text{if } (z, w) \in R_1 \cap R_2 \\ &= 0 && \text{elsewhere,} \end{aligned}$$

where $K_1 + K_2 = 1$. Then $r(Z, W) = r$ if and only if $K = H$.

For example, with the uniforms (inside an ellipse), we can construct the model in Figure 1. The value of r will be preserved as long as the two ellipses have the same center and the same orientation.

Schilling's second observation presents an alternative formula:

$$\hat{r} = [1 - (d/D)^2]/[1 + (d/D)^2].$$

I have applied it on 139 elliptical sample plots taken in manuals, and the precision seems roughly the same as for

$$F = \sqrt{1 - (h/H)^2}.$$

It is perhaps a bit better, but not significantly.

Advantage of Schilling's Formula. With a rule, it is effectively more natural and easy to measure the principal axes than the vertical heights.