Striking Demonstrations in Teaching Statistics

Eric R. Sowey
The University of New South Wales

Journal of Statistics Education Volume 9, Number 1 (2001)

Copyright © 2001 by Eric R. Sowey, all rights reserved.
This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the author and advance notification of the editor.

Key Words: Enrichment materials for teaching; Intellectual excitement; Long-term learning.

Abstract

Long-term learning should, surely, be an outcome of higher education. What is less obvious is how to teach so that this goal is achieved. In this paper, one constructive contribution to such a goal is described in the context of statistical education: the introduction of striking demonstrations. A striking demonstration is any proposition, exposition, proof, analogy, illustration, or application that (a) is sufficiently clear and self-contained to be immediately grasped, (b) is immediately enlightening, though it may be surprising, (c) arouses curiosity and/or provokes reflection, and (d) is so presented as to enhance the impact of the foregoing three characteristics. Some 30 striking demonstrations are described and classified by statistical subfield. The intent is to display the variety of devices that can serve effectively for the purpose, as a stimulus to the reader's own enlargement of the list for his or her own pedagogical use.

1. Introduction

Where is the excitement in teaching and in learning? This question tantalises academics and their students everywhere, though the search for an answer in practice is often neglected.

In a previous paper (Sowey 1995), I have argued that intellectual excitement grows from teaching where (i) students see the discipline as one of central importance, but one in which not everything is yet settled, (ii) the teacher's enthusiasm for and commitment to the discipline is evident, and (iii) some striking demonstrations are introduced that will arouse students' curiosity and/or provoke reflection. In this paper I want to explore further the identification and use of striking demonstrations in teaching statistics.

2. What Is a Striking Demonstration?

I shall use "demonstration" here in a very broad sense to mean any proposition, exposition, proof, analogy, illustration, or application presented to students. The presentation may be such that the teacher reveals all, or it may involve various degrees of student interaction and discovery. Further, it may or may not be computer-mediated. The quality of presentation is as important as the significance of content in making effective educational use of demonstrations -- indeed, some would say, more important.

A demonstration is striking when

a. it is sufficiently clear and self-contained to be immediately grasped.
b. it is immediately enlightening, though it may be surprising.
c. it arouses curiosity and/or provokes reflection.
d. it is so presented as to enhance the impact of the foregoing three characteristics.

In the evocative language of Martin Gardner (1982), a striking demonstration produces an "aha!" reaction: "Aha! now I
understand," "Aha! what a great way to get that result," "Aha! I see. But that contradicts what I already know. Where's the catch?"

A striking demonstration in statistics can take many forms. To cite a few possibilities: it may appear as a unifying formulation of at-first-sight diverse results, as a counterintuitive (but true) proposition, as a logical paradox, as a counterexample to a seemingly general principle, as an analogy with something already familiar to the student, as a vivid geometric, numeric, or graphical illustration of some abstract principle or algebraic theorem, or as an unusual and attention-gripping application of familiar statistical tools. Moreover, it need not be restricted to ideas that "work"; displaying ideas that "don't work" can be highly instructive in its own way (see, below, Examples 1 and 2 in Section 4.4, and Examples 1 and 3 in Section 4.6). The mode of presentation can enhance the effect of such demonstrations. The current emergence of interactive demonstrations over the World Wide Web offers particularly attractive presentation possibilities.

3. Why Use Striking Demonstrations?

The immediate answer is "to help make the subject memorable." A striking demonstration is not just an end in itself; it is also a hook on which to hang an exposition of related, but less immediately appealing, material. The stimulus of the former may then be expected to catalyse study effort on the latter. Exploiting this "halo effect" is an effective way to engender learning that lasts -- beyond the semester assessments, beyond the academic year, and, indeed, beyond the end of the degree program. Lasting knowledge should, surely, be a goal of all university education!

In some disciplines the use of striking demonstrations has a long and distinguished history. Physics has been a leader in this regard. Galileo's demonstration from the Leaning Tower of Pisa in 1590 that a body's acceleration under gravity is independent of its mass, and Michael Faraday's demonstration of electrical induction before the Royal Society in 1831 are two prominent examples. This tradition continues in university teaching of the physical sciences today (see Taylor 1988). Statistics, too, has a tradition of lecture-demonstrations with physical apparatus, from Galton's quincunx in 1874 (see Stigler 1986, pp. 275-281) up to the present day (for example, Loosen 1997).

But the educational significance of striking demonstrations as a stimulus to long-term learning of the broader subject matter has received only incidental recognition in statistics textbooks, and -- probably for that reason -- in teaching approaches, as well. This does not, of course, mean that there exist few striking demonstrations in statistics. Rather, it means that very little systematic effort has been made to identify such demonstrations for use in teaching. This may be changing, particularly with regard to interactive demonstrations as the World Wide Web enriches channels of communication. But most of the effort devoted to developing and collecting such demonstrations serves the need of foundation courses only. The presentation of advanced courses in statistics is still generally as free of striking demonstrations today as it always was.

4. Striking Demonstrations in Teaching Statistics

A particular few striking demonstrations are widely used in teaching statistics (including its probability foundations) -- so widely, indeed, that one might even call them hackneyed examples of the genre. Two such are the "birthday problem" (see, for example, Mosteller, Rourke, and Thomas 1970, pp. 97-98) and a demonstration of the central limit theorem effect via simulation. It is puzzling why just these few instances of striking demonstrations should have come into routine pedagogical use, when so much more can be done.

In what follows, I offer some 30 striking statistical demonstrations that I have found, from student feedback over many years, to be highly effective in my own teaching. In most cases I include a source in the literature, where one is known to me, where a fuller discussion can be found. My aim is to display the variety of devices that can serve as striking demonstrations, as a stimulus to the reader's own enlargement of the list for his or her own pedagogical use. The examples given are not keyed solely to an introductory level of exposition. For ease of reference, the demonstrations are classified according to statistical subfield, and each is prefaced by a concise description of its message.

4.1 Descriptive Statistics

1. Selective choice of an average can misrepresent a real-world situation.

Consider two equally priced goods. Between two years, the price of one doubles (from 100 to 200, say), and the price of the other halves (from 100 to 50). Now, what would you like to show? Average price up? Then use the arithmetic mean of 200 and 50. Average price down? Then use the harmonic mean. Or no change? Then use the geometric mean.

Source: Huff (1954), p. 118.
2. In some real-world contexts the mean is inappropriate as a measure of central tendency.

"Most people in this city have more than the average number of legs."
Name an average whose calculation will lead to this conclusion.

4.2 Elementary Probability

1. If an event $A$ is impossible, then $P(A) = 0$, but the converse is not necessarily true.

Consider, as an example, $P(X = k, a constant)$ in the distribution of any continuous variable $X$ whose domain includes $k$. Or, alternatively, suppose a coin is tossed. Then there are (at least) four possible outcomes: head, tail, edge, and swallowed by a passing bird. To the last two possibilities, we customarily (implicitly!) assign zero probability. Are these impossible events? Perhaps we should rather call them "infinitely unlikely."

2. The formal meaning of independence in statistics and its intuitive meaning in daily life are not always in harmony.

Two coins are tossed at random. Consider the events:

- $A$: A head occurs on the first coin
- $B$: The coins fall alike

Then $A$ and $B$ are statistically independent, since $P(AB) = P(A)P(B)$, but intuitively the events $A$ and $B$ are systematically related.

An extension of this idea produces Bernstein's Paradox, namely, that pairwise independence of a set of $n$ events does not assure overall independence of the events.

Source: Szekely (1986), pp. 12-16.

3. Aggregation of data can produce counterintuitive effects on probabilities.

The data in the following three tables illustrate the phenomenon known as Simpson's Paradox.

Table 1. Contingency Table for Males Treated/Not Treated With a New Drug

|          | Treated | Not treated |
|----------|---------|-------------|
| Recovered| 700     | 80          |
| Not recovered| 800     | 130         |

Table 2. Contingency Table for Females Treated/Not Treated With a New Drug

|          | Treated | Not treated |
|----------|---------|-------------|
| Recovered| 150     | 400         |
| Not recovered| 70      | 280         |

Table 3. Contingency Table (Aggregated) for Persons Treated/Not Treated With a New Drug

|          | Treated | Not treated |
|----------|---------|-------------|
| Recovered| 850     | 480         |
| Not recovered| 870     | 410         |
Assign symbols to events as follows:

- A  "Recovered"  A*
- B  "Treated"  B*
- C  "Male"  C*
- A  "Not recovered"
- B  "Not treated"
- C  "Female"

Then from Table 1, \( P(A|BC) = 700/1500 = 0.47 \) and \( P(A|B^*C) = 80/210 = 0.38 \).

Conclusion: Treatment is more effective than no treatment.

From Table 2, \( P(A|BC^*) = 150/220 = 0.68 \) and \( P(A|B^*C^*) = 400/680 = 0.59 \).

Conclusion: Treatment is more effective than no treatment.

But from Table 3, \( P(A|B) = 850/1720 = 0.49 \) and \( P(A|B^*) = 480/890 = 0.54 \).

Conclusion: Treatment is less effective than no treatment!

Sources: Szekely (1986), pp. 58 and 135-136, Wagner (1982).

4.3 Statistical Distributions

1. Poisson probabilities of time-dependent events are not constant under proportionate change in both number of occurrences and time interval.

Suppose a variable \( X \) has a Poisson distribution with mean, say, \( \mu \) occurrences per hour. Then one might intuitively expect that the probability of \( x \) occurrences in one hour would be the same as \( 2x \) occurrences in two hours. But this is not the case, for the variance of \( X \) is also \( \mu \), and hence grows with the increase in mean number of occurrences in the longer time period.

2. The tails of the normal distribution lie much closer to the horizontal axis than conventional textbook diagrams depict.

If the standard normal distribution is drawn to scale on a sheet of paper, so that its tails are 1 millimetre above the axis at \( z = 6 \), how high must the paper be at the mode to accommodate the drawing? The surprising answer is 65.7 kilometres (see Sowey 1995, Section 4.1).

This example shows just what is meant by a "thin-tailed distribution" in statistics. By contrast, the \( t \)-distribution is very much a "fat-tailed distribution." At 10 degrees of freedom, the \( t \)-variate's ordinate at \( t = 6 \) is 14864 times higher than the \( z \) ordinate at \( z = 6 \). At 20 degrees of freedom, the corresponding multiple is 1325; at 30 degrees of freedom it is 323.

3. A simple probability density function may have a limited number of finite moments.

It surprises many students, shown a plot of the Cauchy distribution with an evident centre, to be told that the mean of this distribution is infinite. A probability density function (pdf) of simple form that can be used to clarify this notion is

\[
f(X) = \alpha/(X + \alpha)^2 \quad X \geq 0, \alpha \neq 0
\]

This function, like the Cauchy pdf, has no finite moments, but it is a more useful example than the Cauchy because its form is flexible. Thus, a simple variant

\[
f(X) = r\alpha^r/(X + \alpha)^{r+1} \quad X \geq 0, \alpha \neq 0
\]

has finite moments only up to order \( (r - 1) \).

Source: Lachenbruch and Brogan (1971).
4.4 Statistical Estimation

1. Patterns are not always generalisable in the choice of optimal estimator.

   In a foundation statistics course, students learn that
   - for the normal distribution, the "best" estimator of the population mean is the sample mean;
   - for the normal distribution, the "best" estimator of the population variance is the sample variance;
   - for the binomial distribution, the "best" estimator of the population proportion is the sample proportion.

   Students are tempted to conclude that this pattern is generalisable, i.e., the "best" estimator for a particular population parameter is always the corresponding sample statistic. (Goldberger 1991, p. 117, calls this "the analogy principle.") However, such a conclusion would be mistaken.

   Counterexample:
   For the rectangular distribution, the "best" estimator of the population mean is the sample midrange.

   Source: Romano and Siegel (1986), p.189.

2. Though there may be a well-defined routine procedure for generating a class of estimators, this does not mean it will invariably be operational.

   a. Maximum likelihood estimation (MLE).

      i. The likelihood surface may be totally flat at its maximum, so that the MLE is not unique. This is found to be the case, for example, in seeking the MLE of a parameter $\theta$ in the rectangular distribution $f(X) = 1$, for $\frac{\theta}{2} \leq X \leq \theta + \frac{\theta}{2}$.

      Source: Romano and Siegel (1986), p. 182.

      ii. The likelihood function may be unbounded, so that there is no maximum of the likelihood.

      Example 1: The errors-in-variables model of simple regression, where the ratio of error variances is not an a priori fixed quantity.

      Sources: Johnston (1963), p. 152, and Kmenta (1972), p. 309.

      Example 2: The regression model with ongoing randomly missing measurements, where each missing measurement is estimated as an individual parameter along with the regression coefficients.

      Source: Kmenta (1986), p. 382.

   b. Minimum mean square error (MSE) estimation.

      Examples: The minimum mean square error estimator of $\mu$ in $N(\mu, \sigma^2)$ involves $\mu$, and hence is not a useful estimator. Similarly for the minimum MSE estimator of $\beta$ in the multiple regression model

      $Y = X\beta + \epsilon$.

      Source: Kmenta (1986), pp. 186, 219.

   c. Any estimator in the context of an undersized sample.

      It is not always appreciated that the definition of an undersized sample differs according to the estimation context, and, in particular, differs between single equation and simultaneous equation models.

      Source: Klein (1973).

      However, analogy is not always a reliable guide. For instance, unique parameter estimation is impossible in a single-equation least squares regression if the number of predetermined variables ('regressors') exceeds the number of observations. Yet, two-stage least squares estimation is possible in a simultaneous equation system even if the total number of predetermined variables in the system exceeds the number of observations.

      Source: Fisher and Wadveki (1971).
3. An unbiased estimator that is optimal on the minimum variance criterion turns out suboptimal on the minimum mean square error criterion, and is, in fact, dominated by a biased estimator.

Consider $k$ independent $N(\mu_i,1)$ ($i = 1, 2, 3, ... , k$) distributions, with $X_i \sim N(\mu_i,1)$. Then the maximum likelihood estimator of the vector of means, $(\mu_1, \mu_2, \mu_3, ...)$, based on a sample of size one from each distribution, is $(X_1, X_2, X_3, ...)$. In 1956, Charles Stein showed that, provided that $k \geq 3$, using the estimator $[1 - (1/s^2)]X_i$ ($i = 1, 2, 3, ...$), where $s^2 = \sum_{i=1}^{k} X_i^2$, instead of the estimator $X_i$ will produce an estimated mean vector with an unambiguously smaller expected mean square error. This remarkable counterintuitive result is known as Stein's paradox.

Sources: Efron and Morris (1977); Thompson (1989), pp. 177-183; Stigler (1990).

4.5 Statistical Testing

1. The "power" of a significance test is conceptually (though not formally) related to the "power" of an optical lens.

This illustrates the possibilities opened up by using an analogy with something already familiar to the student. Just as the unaided eye (a low-power lens) cannot reliably distinguish whether a far distant figure is a single person or two close together, neither can a statistical test of low power reliably distinguish between the null hypothesised and true values of a parameter if those values are close together.

2. The results of significance tests in statistical models are not invariant under observationally-equivalent model respecifications.

Example: Seasonal regressions using dummy variables.

Whichever three of the four seasonal dummies are included in the regression model, the resulting models are observationally equivalent. However, when working with sample data, while coefficients obtained using, say, the March, June, and September seasonal dummies might all be statistically significant, this does not imply that all the coefficients would necessarily also be significant if the June, September, and December dummies were used instead. Thus, forecasts obtained from alternative specifications of the same underlying regression model may differ.

The following illustration can help make this clear. The dataset represents quarterly observations of Australian Gross Farm Product (GFP), in millions of current dollars, over seven successive years.

| Year | March | June | September | December |
|------|-------|------|-----------|----------|
| 1    | 1082  | 305  | 579       | 1900     |
| 2    | 892   | 394  | 805       | 1709     |
| 3    | 1179  | 498  | 875       | 1652     |
| 4    | 836   | 570  | 976       | 2309     |
| 5    | 2098  | 1092 | 1426      | 3354     |
| 6    | 1627  | 988  | 1574      | 3225     |
| 7    | 1383  | 953  | 1466      | 3754     |

a. Least squares regression of GFP on a time trend, $T$ ($= 1, 2, 3, ... , 28$), and March, June, and September intercept dummy variables:

\[
\text{GFPpred} = 1769.6 + 49.25T - 1110.3D_m - 1773.4D_j - 1408.2D_s
\]

\[
(9.34) (5.88) (-5.81) (-9.32) (-7.43)
\]

"GFPpred" is the regression predicted value of GFP.

In this estimated regression, all the coefficients are statistically significant.

The implied individual-quarter regression models are:
b. Least squares regression of GFP on a time trend, $T$, and June, September, and December intercept dummy variables:

\[
\begin{align*}
\text{GFP}_{\text{pred}} &= 659.3 + 49.25T - 663.1D_j - 297.9D_s + 1110.3D_d \\
&= 526.5 + 48.10T - 514.1D_j + 1261.5D_d
\end{align*}
\]

(t-statistics with 23 degrees of freedom in parentheses)

The regression coefficient on the September dummy is here statistically insignificant.

If all dummy variables in this regression are, nevertheless, retained in the model, this regression implies the same individual-quarter regression models as in part (a).

But if, as is statistically appropriate, the insignificant regressor is eliminated, and the resulting model is re-estimated by least squares, then the following estimation results:

\[
\begin{align*}
\text{GFP}_{\text{pred}} &= 526.5 + 48.10T - 514.1D_j + 1261.5D_d \\
&= 526.5 + 48.10T
\end{align*}
\]

(t-statistics with 24 degrees of freedom in parentheses)

This will produce predictions different from those in part (a), since it implies the following individual-quarter regression models:

| Month   | GFPpred     |
|---------|-------------|
| March   | $659.3 + 49.25T$ |
| June    | $-3.8 + 49.25T$  |
| September | $361.4 + 49.25T$ |
| December| $1769.6 + 49.25T$ |

4.6 Regression and Correlation

1. Attempting to define a "line of best fit" by setting to zero the algebraic sum of residuals of scatter points from the fitted line does not work.

This criterion represents only one restriction on the slope and intercept coefficients of the fitted line, but two independent restrictions are needed to obtain unique values of the fitted line's two parameters. So this method will not work, however intuitively plausible it may seem to a statistical novice. The least squares criterion, by contrast, does provide two independent restrictions.

2. What shape of scatter diagram has an $r^2$ value of 0?

If all the points in a scatter lie exactly on the least squares regression line, then $r^2 = 1$. Does this help us discover a shape for a scatter diagram for which $r^2 = 0$? Yes, if we read the first result as systematic movement of $Y$ with $X$ in a particular direction, and ask what would be a shape for a scatter in which $Y$ did not move systematically with $X$ in any direction?

3. Before evaluating any statistical measure, always be sure it is appropriate for the context.

Mark on co-ordinate axes the points (0,2), (2,2), (4,2) and (6,2). Consider these points as a bivariate scatter diagram. What is the value of $r^2$ for this scatter? Is it 1, "because there is an obvious 'line of best fit' on which all the scatter points lie"? Is it 0, "because the regression line will surely be horizontal"?

4. When reliable theory to underpin model-building is lacking in any disciplinary context, always look first at the pattern of the data before choosing a statistical model.
For the following data, perform these four simple regressions:

(a) \( Y_1 \) on \( X_1 \)  
(b) \( Y_2 \) on \( X_1 \)  
(c) \( Y_3 \) on \( X_1 \)  
(d) \( Y_4 \) on \( X_2 \)

|   | \( X_1 \) | \( X_2 \) | \( Y_1 \) | \( Y_2 \) | \( Y_3 \) | \( Y_4 \) |
|---|---|---|---|---|---|---|
| 1 |  8 |  8 |  8 |  8 |  7 |  5 |
| 2 | 10 | 13 |  5 |  8 |  8 |  8 |
| 3 | 11 |  9 |  9 |  8 |  8 | 10 |
| 4 |  6 | 11 |  7 |  8 |  8 |  9 |
| 5 |  7 |  7 |  8 |  8 |  8 |  7 |
| 6 |  4 |  8 |  9 |  8 |  8 |  5 |
| 7 |  7 |  6 |  7 |  8 |  8 |  5 |
| 8 |  8 |  8 |  8 |  8 |  8 |  5 |
| 9 |  8 | 10 |  8 |  8 |  8 |  5 |
| 10|  8 |  8 |  8 |  8 |  8 |  5 |

All these regressions produce the same estimated values (correct to two decimal places) for the intercept (3.00) and the slope (0.50). The value of \( R^2 \) is also the same in each case (0.67). When the data are plotted, however, each set tells its own very different story. The folly of computing blindly is powerfully demonstrated.

*Source: Anscombe (1973).*

5. Evidence of collinearity: \( R^2 \) and the \( F \)-statistic are high, yet all \( t \)-statistics are insignificant.

These characteristics can be easily illustrated with this dataset (for which \( X_2 + X_3 = 4 \)), when \( Y \) is regressed on \( X_2 \) and \( X_3 \):

|   | \( Y \) | \( X_2 \) | \( X_3 \) |
|---|---|---|---|
| 1 |  -3 |  -1 |   0   |
| 2 |  -1 |   0 |   2   |
| 3 |   5 |   4 |   2   |
| 4 |   3 |   6 |   1   |

\( R^2 = 0.97, F = 29.8, t \)-ratio for \( \beta_2 = 1.26, t \)-ratio for \( \beta_3 = -0.68. \)

*Source: Geary and Leser (1968).*

6. A problem stated algebraically is sometimes immediately elucidated when viewed geometrically.

a. The property \( \text{var}(X + c) = \text{var}(X) \), where \( c \) is a constant, is seen immediately to hold from a diagram in which the distribution of \( X \) is translated sideways by a distance \( c \).

b. The reason that a least squares regression on perfectly collinear regressors cannot be performed is seen geometrically, in a three-dimensional case, to result from the data scatter being reduced to two dimensions, so that there is not enough information to fix the position of the regression plane uniquely in three-dimensional space.

c. When a regression model includes among the regressors a two-phase (0,1) dummy variable that is "switched on" for only a single observation, least squares estimation of the model produces the same regression coefficients, for the regime in which the dummy variable is "switched off," as simply deleting the observation for which the dummy is "switched on" and recomputing the regression, but this time omitting the dummy variable as a regressor.

This result is transparent when represented geometrically in a diagram whose construction is now described.

Sketch, in perspective, a Cartesian axes system with axes labelled \( X, D, \) and \( Y \). The \( Y \) axis points "upwards," and the \( X \) and \( D \) axes between them define the plane of the "floor."

On these axes we represent a three-dimensional scatter diagram for data generated according to the regression model.
Here, $X$ and $Y$ are quantitative variables, $D$ is a $(0,1)$ dummy variable, and $\epsilon$ is a random disturbance. There are $n$ observations. For $n - 1$ of these, $D = 0$ (dummy "switched off"), while for just one observation, $D = 1$ (dummy "switched on").

Along the $D$ axis, mark the points $D = 0$ and $D = 1$. Then at $D = 0$ draw in a "vertical" plane, parallel to the plane defined by the $X$ and $Y$ axes. Call this plane $A$. Next, at $D = 1$ draw in a "vertical" plane, parallel to the plane just drawn at $D = 0$. Call this plane $B$.

If there is just a single scatter point corresponding to the value $D = 1$, then that point will lie on plane $B$. All the remaining scatter points, though distributed widely in $(X, Y)$ space, will be coplanar on plane $A$.

Now, a least squares regression plane for the entire scatter will intersect plane $A$ along a line of best fit for those scatter points that are coplanar on plane $A$. What will be the slope of this regression plane in $(X, D, Y)$ space? Well, there is only one other scatter point in the full co-ordinate space, and that is the point that lies on plane $B$. Accordingly, the regression plane will head directly towards that point and pass through it. The consequence will be that the regression residual corresponding to that point will necessarily be zero.

It follows that the single scatter point on plane $B$ adds nothing to the sum of squared residuals, whose minimisation defines the regression estimates of the parameters $\beta_0$, $\beta_1$, and $\beta_2$. Thus, the estimates of the regression coefficients, $\beta_0$ and $\beta_1$, for the regression model estimated with the dummy variable included, and specialised to the regime where the dummy variable is "switched off" (thereby excluding $\beta_2$), will be the same as would be obtained if the scatter point on plane $B$ were deleted and the regression run on $(n - 1)$ observations without the dummy variable.

7. The restricted least squares (RLS) estimator has some paradoxical characteristics.
   a. The Gauss-Markov Theorem assures us that the ordinary least squares estimator in the classical multiple regression model is the best linear unbiased estimator, yet the RLS estimator (also a linear estimator) is more efficient. Where is the catch?  
      *Source:* Goldberger (1964), p. 257.
   b. The RLS estimator is more efficient than the ordinary least squares estimator *whether or not* the exact linear restrictions underlying the RLS estimator are correct. Where is the catch?  
      *Source:* Fomby, Hill, and Johnson (1988), p. 85.

4.7 Statistical Computation

1. Even a simply written calculation can generate a huge rounding error.

   On a hand calculator, raise $\pi$ to the power $1/(2^{**38})$. In many calculators the answer 'one' will appear, but this must be nonsense, for it would imply that one raised to the power $(2^{**38})$ equals $\pi$.

2. A transparent example of solution instability in ill-conditioned simultaneous equations.

   Solve $x - y = 1$
   $x - 1.00001y = 0$

   Solve $x - y = 1$
   $x - 0.99999y = 0$

   In the second system there has been a minute perturbation in one coefficient. How does this affect the solutions? A graphical exploration immediately reveals all.
4.8 Further Striking Applications of Statistical Tools

The following papers offer intriguing insights on the theory and/or practice of the techniques used in each case.

1. Does age affect master chess? See Draper (1963).
2. The lengths of surnames. See Healy (1968).
3. Statistical estimates of the speed of light. See Youden (1972).
4. How many people pay their fares? See Jagers (1973).
5. How many words did Shakespeare know? See Efron and Thisted (1976).
6. Smoking and lung cancer. See Burch (1978).
7. Survival of book titles on the bestseller list. See Grove (1991).
8. Attributing authorship through statistical analysis of literary style. See Elliott and Valenza (1991).
9. Three tantalising probability paradoxes. See Loyer (1986).
10. Interactive demonstrations via the World Wide Web. See West and Ogden (1998).

5. Presenting a Striking Demonstration

The striking effect of a demonstration, and its consequent "halo" effect on student learning, can be lost if it is indifferently presented. To guard against this, the way it is used in teaching should be carefully premeditated. This recommendation applies to the choice of occasion for introducing it, to the choice of wording used, and to how much of an active role students are given in the presentation. It applies also, and particularly, to the way computer-based interactive demonstrations are introduced -- especially those that are Web-based -- to ensure that the novelty for students of the medium does not mask the message.

Let me take two examples from among the demonstrations given earlier in this paper.

In Section 4.2, Example 1 -- considering the set of outcomes when a coin is tossed -- is very relevant when explaining the inadequacy of Bernoulli's Principle of Insufficient Reason as a general rule for assigning probabilities to events. After this Principle has been discussed in the context of heads and tails, class members should be invited to say what further possibilities there can be when a coin is tossed. "Landing on an edge" is usually volunteered quite quickly, though doubtfully because it seems so unrealistic. To dispel the doubt, explain that the coin is being tossed in a muddy field. Once this is clearly established as a practical possibility, ask for yet further possibilities. If there is silence, tease students by hinting that they need to consider every logical possibility, and so far only possibilities involving the fall of the coin have been examined. This will, sooner or later, elicit the response "the coin doesn't come down," and often the speaker's tone of voice will suggest that he/she cannot quite believe what he/she is saying. This is the moment to draw attention to the passing bird! Experience shows that, discovered in this way, the expanded set of outcomes will remain indelibly in students' minds. Further, the consequence for Bernoulli's Principle of the fact that all the outcomes when a coin is tossed are not equally likely, will be understanding firmly grounded.

In Section 4.3, Example 3 -- illustrating the notion that some of the moments of a distribution may not be finite -- is relevant to conventional ways of summarising and interpreting results from Monte Carlo experiments on the properties of statistical estimators. If an estimator has an infinite mean, then the (finite) Monte Carlo estimator of that mean will be spurious. Similarly, if an estimator has an infinite variance (though a finite mean). It comes as a surprise to many students that the exact sampling distribution of an estimator can have some moments finite and some infinite. The example given allows them to grasp the reality in a simplified setting.

In this context, it is not essential that students, themselves, derive the moments. The valuable thing is that students see, in concrete terms that they can comprehend, an important way in which Monte Carlo analyses are methodologically vulnerable -- an issue that is little remarked on in textbooks.

6. Finding Striking Demonstrations in Statistics

It will be clear from the nature of the preceding examples that striking demonstrations in the scholarly literature do not always announce themselves. They must be recognised as such! Moreover, Web-based interactions, however visually impressive, are not for that reason necessarily educationally enhancing -- a point that tends to get glossed over in the enthusiastic rhetoric of "the new technologies." Creating, refining, and polishing a striking demonstration is a demanding but pedagogically rewarding endeavour. Fortunately, there are also some well-focused and selective collections of material to explore.
a. There are several books of counterexamples and paradoxes in statistics:

Romano and Siegel (1986), Szekely (1986), Stoyanov (1987), and Wise and Hall (1993), but it should be noted that the last two of these are written at an advanced level. Stoyanov (1988) makes proposals for including counterexamples in teaching statistics and probability.

b. There are books, journals, and websites that feature remarkable facets of statistics and probability, and arresting applications of statistical tools in everyday life. Here is a selection.

Books:
Wallis and Roberts (1956), Campbell (1974), Hollander and Proschan (1984), Jaffe and Spirer (1986), Tanur et al. (1989), Wang (1993), Isaac (1995), and Moore (1997).

Journals:
The Journal of Statistics Education. American Statistical Association, vol. 1 = 1993. [Online at http://ww2.amstat.org/publications/jse]
The American Statistician. American Statistical Association, vol. 1 = 1947.
Chance: New Directions For Statistics and Computing. Springer, vol. 1 = 1988.

Apart from these sources, it is good for the teacher to bear in mind that material for a striking demonstration may turn up in any research paper in statistics. Wide reading, while developing a sense for what will serve well for the purpose, is always rewarding.

In the search for striking demonstrations, it can help to first formulate a series of principles that students should carry away with them from study of a subject, for example "never discard an outlier without examination," "avoid 'forcing the fit' in regression modelling," "if a problem is difficult when posed algebraically, try looking at it geometrically." With these principles firmly in mind, it will be easier to recognise striking demonstrations of them when one comes upon them, whatever the context. As Louis Pasteur said, "chance favours the prepared mind."

Acknowledgments

For stimulating my thoughts on the subject of this paper I want to thank (in alphabetical order): Glenys Bishop, Gary Grunwald, Pam Hollis, Paul Lochert, Helen MacGillivray, Meei Ng, Peter Petocz, Ken Sharpe, Pamela Shaw, Bruce Stephens, and Ross Taplin.

My thanks go also to Jan Kmenta for valuable discussions on several aspects, and to the referees of this paper for their very constructive comments. I am grateful to Jackie Dietz for her meticulous editing in preparing this paper for publication.

References

Anscombe, F. J. (1973), "Graphs in Statistical Analysis," American Statistician, 27, 17-21.

Burch, P. R. (1978), "Smoking and Lung Cancer: The Problem of Inferring Cause," Journal of the Royal Statistical Society, Ser. A, 141, 437-477.

Campbell, S. K. (1974), Flaws and Fallacies in Statistical Thinking, Englewood Cliffs, NJ: Prentice-Hall.

Draper, N. R. (1963), "Does Age Affect Master Chess?" Journal of the Royal Statistical Society, Ser. A, 126, 120-127.

Efron, B., and Morris, C. (1977), "Stein's Paradox in Statistics," Scientific American, 236, May, 119-127.
Efron, B., and Thisted, R. (1976), "How Many Words Did Shakespeare Know?" *Biometrika*, 63, 435-447.

Elliott, W., and Valenza, R. (1991), "Who Was Shakespeare?" *Chance*, 4(3), 8-14.

Fisher, W. D., and Wadycki, W. J. (1971), "Estimating a Structural Equation in a Large System," *Econometrica*, 39, 461-465.

Fomby, T. B., Hill, R. C., and Johnson, S. R. (1988), *Advanced Econometric Methods*, New York: Springer.

Gardner, M. (1982), *Aha! Gotcha -- Paradoxes to Puzzle and Delight*, New York: Freeman.

Geary, R. C., and Leser, C. E. (1968), "Significance Tests in Multiple Regression," *American Statistician*, 22, 20-21.

Goldberger, A. S. (1964), *Econometric Theory*, New York: Wiley.

----- (1991), *A Course in Econometrics*, Cambridge, MA: Harvard University Press.

Grove, M. A. (1991), "Survival on the Bestseller List," *Chance*, 4(2), 39-45.

Healy, M. J. R. (1968), "The Lengths of Surnames," *Journal of the Royal Statistical Society*, Ser. A, 131, 567-568.

Hollander, M., and Proschan, F. (1984), *The Statistical Exorcist -- Dispelling Statistics Anxiety*, New York: Dekker.

Huff, D. (1954), *How To Lie With Statistics*, London: Gollancz.

Isaac, R. E. (1995), *The Pleasures of Probability*, New York: Springer.

Jaffe, A. J., and Spierer, H. F. (1986), *Misused Statistics: Straight Talk for Twisted Numbers*, New York: Dekker.

Jagers, P. (1973), "How Many People Pay Their Tram Fares?", *Journal of the American Statistical Association*, 68, 801-804.

Johnston, J. (1963), *Econometric Methods* (1st ed.), New York: McGraw-Hill.

Klein, L. R. (1973), "The Treatment of Undersized Samples in Econometrics," in *Econometric Studies of Macro and Monetary Relations*, eds. A. A. Powell and R. A. Williams, Amsterdam: North Holland, pp. 3-26.

Kmenta, J. (1972), *Elements of Econometrics* (1st ed.), New York: Macmillan.

----- (1986), *Elements of Econometrics* (2nd ed.), New York: Macmillan.

Lachenbruch, P., and Brogan, D. (1971), "Some Distributions on the Positive Real Line Which Have No Moments," *American Statistician*, 25, 46-47.

Loosen, F. (1997), "A Concrete Strategy for Teaching Hypothesis Testing," *American Statistician*, 51, 158-163.

Loyer, M. W. (1986), "Not-so-surprising Surprising Results," in *Proceedings of the Statistical Education Section, American Statistical Association*, pp. 143-147.

Moore, D. S. (1997), *Statistics: Concepts and Controversies* (4th ed.), New York: Freeman.

Mosteller, F., Rourke, R. E., and Thomas, G. B. (1970), *Probability with Statistical Applications* (2nd ed.), Reading, MA: Addison-Wesley.

Romano, J. P., and Siegel, A. F. (1986), *Counterexamples in Probability and Statistics*, Monterey, CA: Wadsworth.

Sowey, E. R. (1995), "Teaching Statistics: Making It Memorable," *Journal of Statistics Education* [Online], 3(2). (ww2.amstat.org/publications/jse/v3n2/sowey.html)

Stigler, S. M. (1986), *The History of Statistics*, Cambridge: Harvard University Press.
----- (1990), "A Galtonian Perspective on Shrinkage Estimators," *Statistical Science*, 5, 147-155.

Stoyanov, J. M. (1987), *Counterexamples in Probability*, New York: Wiley.

----- (1988), "The Use of Counterexamples in Learning Probability and Statistics," in *Proceedings of the Second International Conference on Teaching Statistics*, eds. R. Davidson and J. Swift, The Hague: International Statistical Institute, pp. 280-286.

Szekely, G. J. (1986), *Paradoxes in Probability Theory and Mathematical Statistics*, Dordrecht: Reidel.

Tanur, J. M., Mosteller, F., Kruskal, W. H., Lehmann, E. L., Link, R. F., Pieters, R. S., and Rising, G. R. (eds.) (1989), *Statistics -- A Guide to the Unknown* (3rd ed.), Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Taylor, C. A. (1988), *The Art and Science of Lecture Demonstration*, London: Adam Hilger.

Thompson, J. R. (1989), *Empirical Model Building*, New York: Wiley.

Wagner, C. H. (1982), "Simpson's Paradox in Real Life," *American Statistician*, 36, 46-48.

Wallis, W. A., and Roberts, H. V. (1956), *Statistics -- A New Approach*, Glencoe, IL: Free Press.

Wang, C. (1993), *Sense and Nonsense of Statistical Inference -- Controversy, Misuse, Subtlety*, New York: Dekker.

West, R. W., and Ogden, R. T. (1998), "Interactive Demonstrations for Statistics Education on the World Wide Web," *Journal of Statistics Education* [Online], 6(3). (ww2.amstat.org/publications/jse/v6n3/west.html)

Wise, G. L., and Hall, E. B. (1993), *Counterexamples in Probability and Real Analysis*, Oxford: Oxford University Press.

Youden, W. J. (1972), "Enduring Values," *Technometrics*, 14, 1-11.

---

Eric R. Sowey  
School of Economics  
The University of NSW  
Sydney, NSW, Australia 2052  
E.Sowey@unsw.edu.au

---

**Addendum**

Volume 10, Number 1, of the *Journal of Statistics Education* contains a [Letter to the Editor] concerning this article.

THS, April 1, 2002