PRESIDENTIAL ESSAY NO. 1: APRIL 2021

Notorious: Anscombe's Warning about Diagrams

By Chris Pritchard (chrispritchard2@aol.com)

MATHEMATICAL ASSOCIATION
150th ANNIVERSARY 1871 - 2021
Supporting mathematics in education

Consider the following four data sets:

Set 1		Set 2	
x	*y*	*x*	*y*
10.0	8.04	10.0	9.14
8.0	6.95	8.0	8.14
13.0	7.58	13.0	8.74
9.0	8.81	9.0	8.77
11.0	8.33	11.0	9.26
14.0	9.96	14.0	8.10
6.0	7.24	6.0	6.13
4.0	4.26	4.0	3.10
12.0	10.84	12.0	9.13
7.0	4.82	7.0	7.26
5.0	5.68	5.0	4.74

Set 3		Set 4	
x	*y*	*x*	*y*
10.0	7.46	8.0	6.58
8.0	6.77	8.0	5.76
13.0	12.74	8.0	7.71
9.0	7.11	8.0	8.84
11.0	7.81	8.0	8.47
14.0	8.84	8.0	7.04
6.0	6.08	8.0	5.25
4.0	5.39	19.0	12.50
12.0	8.15	8.0	5.56
7.0	6.42	8.0	7.91
5.0	5.73	8.0	6.89

Rather than arising naturally through measurement (for example), they were crafted in 1973 by the statistician, Frank Anscombe (1918–2001), and as we shall see not only are they quirky but they teach us the importance of the visual in mathematics.

Frank had graduated from Trinity College Cambridge, done some research on rockets during the war and spent a little time at Rothamsted (founded under R A Fisher decades earlier). Later he moved to the States, to Princeton in 1956 and then on to Yale to lead the statistics department. Here he became a leading exponent of the use of computers in statistical analysis. (One of his colleagues at Yale was John Tukey who gave us box plots and other graphical tools; in fact, Anscombe and Tukey married the sisters Phyllis and Elizabeth Rapp.)

The paper in which the data sets appeared is entitled 'Graphs in statistical analysis'. Published in *The American Statistician* and it is online at www.sjsu.edu/faculty/gerstman/StatPrimer/anscombe1973.pdf

Anscombe started by arguing that we pay far too little attention to graphs in statistics, having been 'indoctrinated with these notions', which he lists:

- numerical calculations are exact, but graphs are rough,
- for any particular kind of statistical data, there is just one set of calculations constituting a correct statistical analysis,
- performing intricate calculations is virtuous, whereas actually looking at the data is cheating.

In fact, we should study the output from calculations (the sort of summary statistics we associate with exploratory data analysis) **and** the appropriate associated graphs as 'each will contribute to understanding'.

Anscombe pointed out that graphs 'help us perceive and appreciate some broad features of the data ... and let us look behind those broad features and see what else is there'. Then we come to the crux:

> 'Most kinds of statistical calculation rest on assumptions about the behaviour of the data. These assumptions may be false, and then the calculations may be misleading. We ought always to try to check whether the assumptions are reasonably correct; and if they are wrong we ought to be able to perceive in what way they are wrong. Graphs are very valuable for these purposes.'

He then showed what happens when the summary statistics are calculated and the linear regression model is applied. There are two WOW moments. The summary statistics provide the first, in that they differ not one jot from set to set.

	Set 1	**Set 2**	**Set 3**	**Set 4**
Number of observations	11	11	11	11
Mean of x values	9	9	9	9
Mean of y values	7.5	7.5	7.5	7.5
Sample variance of x	11	11	11	11
Sample variance of y	4.125	4.125	4.125	4.125
Correlation coefficient	0.816	0.816	0.816	0.816
Least squares regression line	$y=\frac{1}{2}x+3$	$y=\frac{1}{2}x+3$	$y=\frac{1}{2}x+3$	$y=\frac{1}{2}x+3$

And here is the second: the graphs (which are taken directly from Anscombe's paper) are completely different from each other.

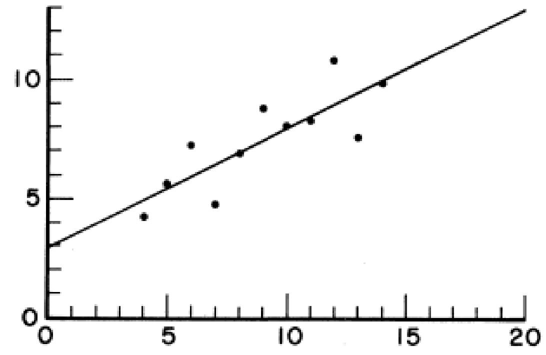

Figure 1

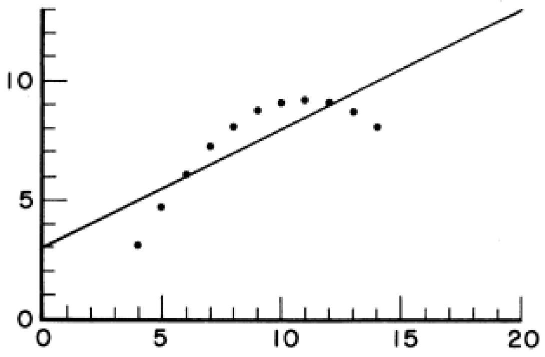

Figure 2

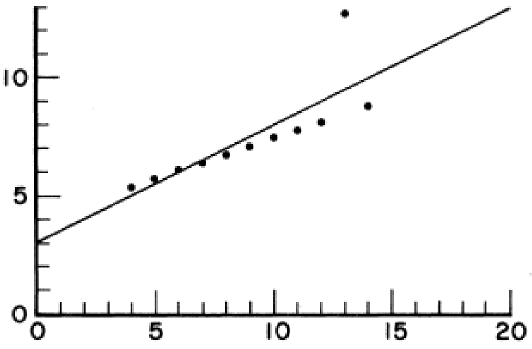

Figure 3

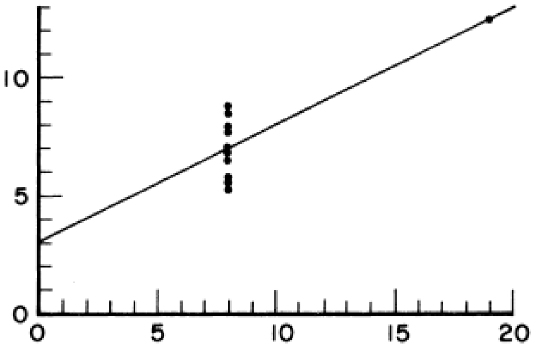

Figure 4

Everything looks right about the first scenario. The scatter plot does suggest that linear regression is wholly appropriate. The array of points in the second suggests non-linear regression, possibly quadratic and the conformity with such a pattern is very strong (little residual variability). The last two have outliers, points that simply do not belong. In Figure 3, that stray point is dragging the gradient of the regression line upwards. Linear regression is appropriate for the other points, giving the equation $y = 0.35x + 4$. It would seem to me (Anscombe didn't quite reach the same view) that the relationship demonstrated in Figure 4 is that regardless of the y value, the associated x value is fixed. So y is the independent variable and it simply doesn't matter what you put into the y machine, the value 8 always pops out. Again, everything is distorted by the outlier.

Of course, the key lesson here is about the dangers of working without a diagram. And of course, this applies to other areas of mathematics, not just statistics.