expected to find that the Poisson distribution described this data. (For the purpose of the experiment, cars were counted travelling in one direction only and "cars" included buses, trucks, etc. but not motor bikes.)

The results of the two experiments are shown in the histograms in Figures 1 and 2 respectively. For example, in the first experiment, no cars were recorded in 30 of the 10 second intervals, 1 car was recorded in 6 of the 10-second intervals, 2 cars were recorded in 19 of the 10-second intervals, 3 cars were recorded in 12 of the 10-second intervals, and so on.

In the first experiment with heavy interrupted traffic (Figure 1), the total number of 10-second intervals recorded is 103 and the average number of cars passing in a 10-second interval is λ = 3.53. With these values, we can calculate the expected number of 10-second intervals during which any given number of cars will pass from the formula (1). For example, the expected number of 10-second intervals with 2 cars recorded is

$$103\, P(2) \;=\; 103\, e^{-3.53}(3.53)^2/2! \;=\; 18.8.$$

These expected values are shown by the broken line in Figure 1. As we anticipated, the expected frequencies do not fit the observed results at all well. This emotional judgment can be confirmed by performing a chi-squared goodness of fit test and this does indeed indicate that the Poisson distribution could not be used to describe this particular type of traffic flow.

The same analysis was done with the data from the second experiment with freely flowing traffic, with the results shown in Figure 2. Again, a subjective judgment and a chi-squared test show that this data also cannot be described by a Poisson model. Apparently, the cars were not arriving independently as we had thought, but were moving in clusters. In fact, the location we used was at a section of the freeway which had traffic lights controlling the flow, but these were obscured from the experimental site by a sharp bend in the road. (This may also serve to confirm a widely held prejudice on the part of the great motoring public that there is no such thing as freely flowing traffic.) The experimental data can be described by the Neyman distribution, a more complicated affair, which copes with situations in which the quantity being observed arrives in clusters, but the various clusters arrive independently.

Can you find any freely flowing traffic?

ARE EXAMINATIONS ANY USE?

M. K. Vagholkar*

If examinations do nothing else, they at least produce vast quantities of numerical data. These numbers are used by all sorts of people, usually with the idea that they indicate a person's knowledge on a particular subject and that they can be used to predict the degree of success in employment or further study. For example, we might suspect that results in the Higher School Certificate examination could be used to predict performance in first year Mathematics at University. How can we test our suspicion? This sort of question is very common in statistics and it has led to many fruitful applications of statistics in practice.

**Dr Vagholkar is a Senior Lecturer in the Department of Statistics at the University of New South Wales.*

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X (Mark in 3-unit Mathematics)	135	95	97	121	104	127	101	108	99	95	92	80	116	86	103	116	104	92	104	80
Y (Mark in first year Mathematics)	75	51	55	70	58	73	56	65	58	58	40	24	75	36	64	58	61	53	49	59

Table 1

Table 1 gives the marks X and Y of 20 students in 3-unit Mathematics in the Higher School Certificate examination and in first year Mathematics at the University. We would expect that if a student does well in 3-unit Mathematics, then he should do well in first year Mathematics, and vice versa. This can be illustrated graphically by the scatter diagram of Y against X in which we plot the points (X,Y) given by the marks in Table 1. (See Figure 1.) There is a mutual dependence between the variables Y and X. Can we find a formula, $Y = f(X)$ say, giving Y in terms of X? An exact formula of this sort is unlikely because, although Y depends on X, it does not depend only on X. There are other factors, both major and minor, besides the mark in 3-unit Mathematics which can influence and decide a student's mark in first year Mathematics. So what we really want is a statistical relationship $Y = f(X) + E$, where $f(X)$ is some function of X and E is a random error variable which measures the contribution made by the factors other than the 3-unit Mathematics mark. We would expect E to be small if these other factors do not make a major contribution to the mark in first year Mathematics and, in that case, our relationship between Y and X would be fairly precise.

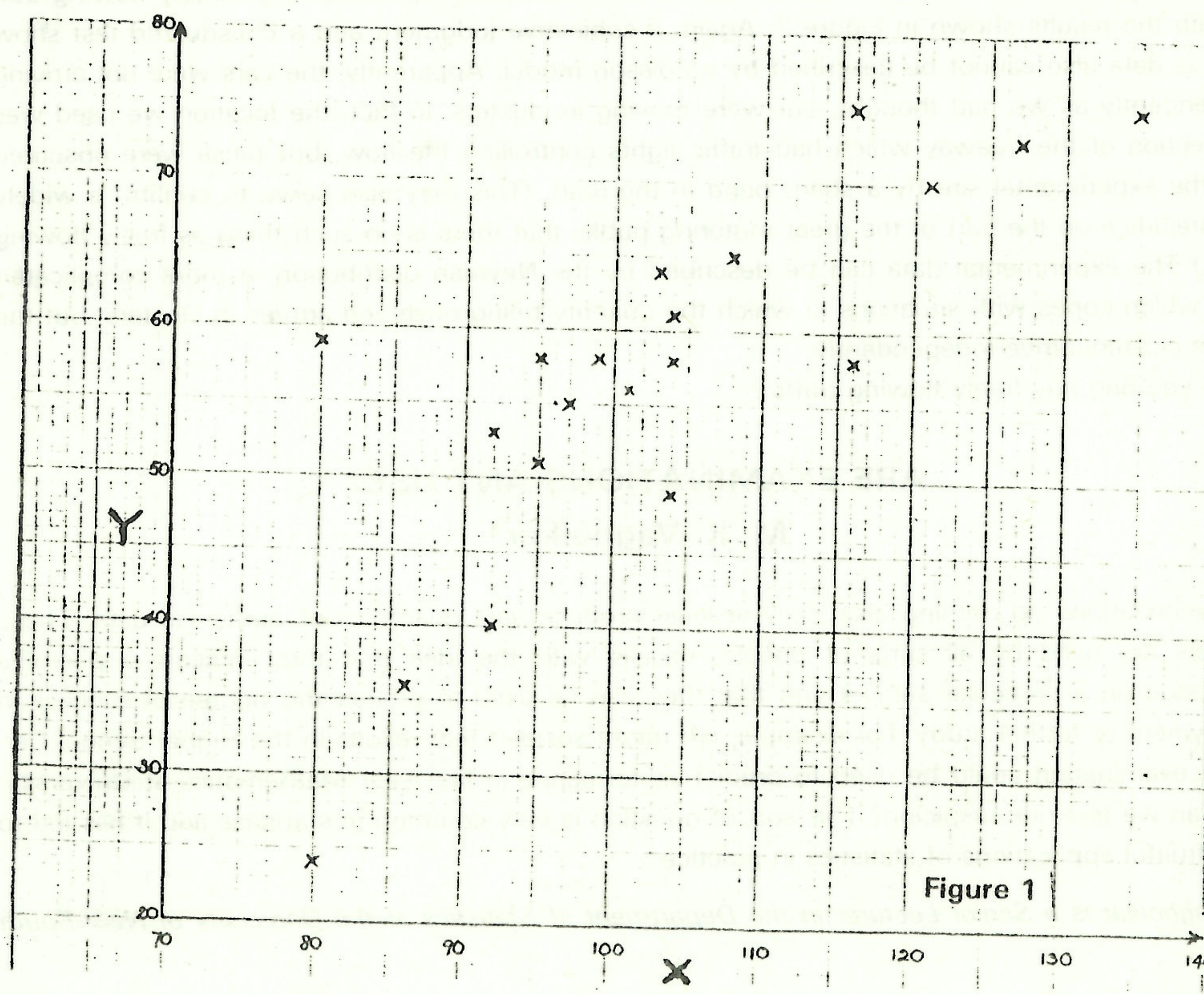

Figure 1

To find an appropriate form for the function f(X), we look at the scatter diagram in Figure 1 and try to guess the curve Y = f(X) around which the points in the diagram cluster. The simplest assumption we can make is that f(X) is a linear function of X, that is the statistical relationship is Y = (a + bX) + E. In other words, we are trying to find the straight line Y = a + bX which gives the best fit to the points in Figure 1. Now we only have to determine the coefficients a and b. This is usually done by the method of least squares, which works as follows. Consider student 16 in Table 1, whose marks are X = 116 and Y = 58. If we use the line Y = a + bX to predict Y from X, the predicted mark for this student is Y = a + 116b. His actual mark is 58, so the error we make is 58 − (a + 116b). The method of least squares determines a and b so that the sum of the squares of these errors for all 20 students is as small as possible. For the data in Table 1, the line of best fit obtained in this way turns out to be Y = −12.67 + 0.68X, as shown in Figure 2. The method of least squares is easy to use, but it does have some drawbacks. Points which lie well off the line have a disproportionately large effect on the sum of squares of the errors and this may lead to undesirable results. Unfortunately, there is no really satisfactory alternative.

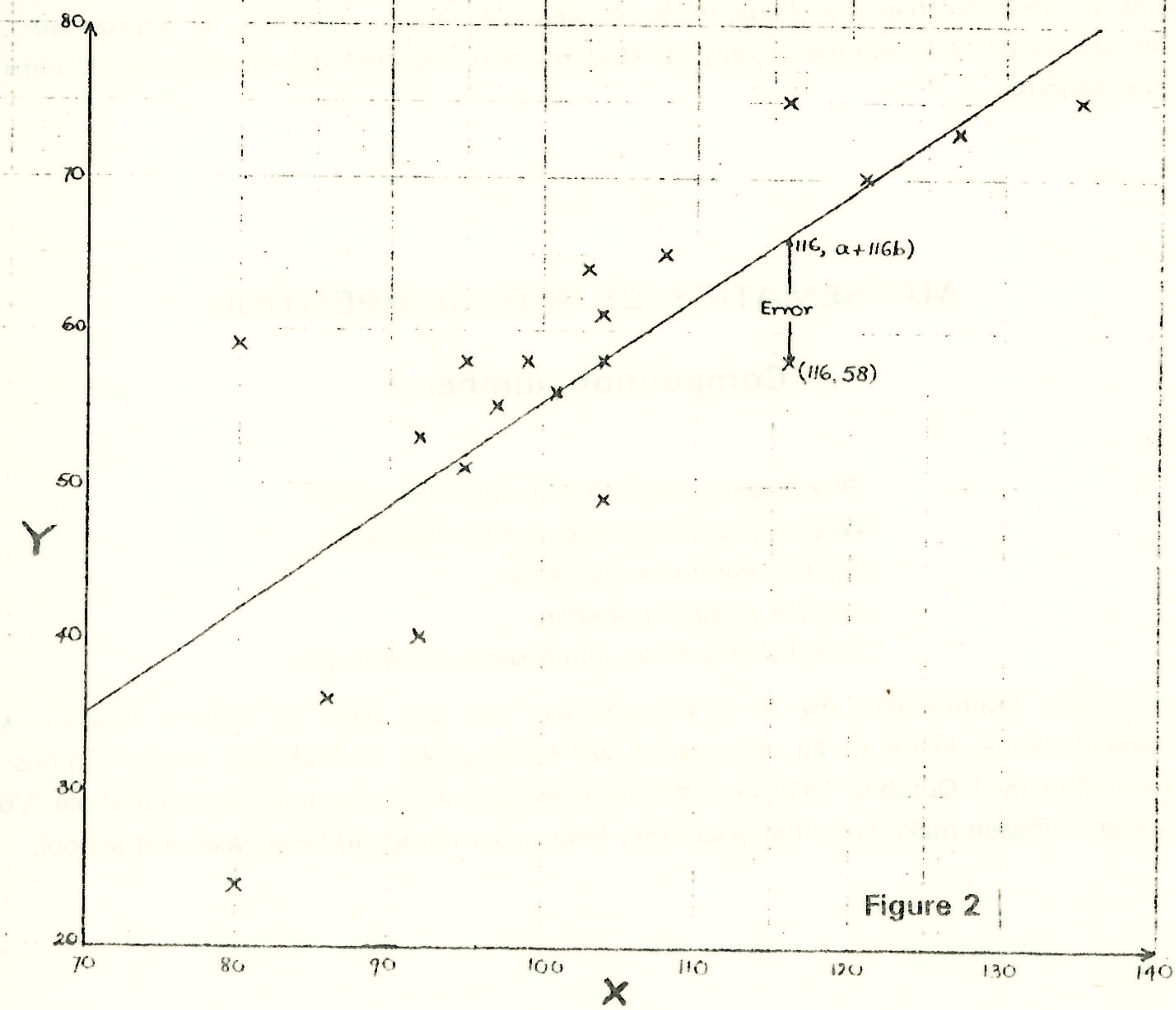

Figure 2

Now that we have found our line of best fit, we must decide how well it fits the data. Various statistical tests are available for this purpose. One way to measure the success of the line of best fit is to calculate the percentage of the total variation in Y-values which is explained by the fitted relationship $Y = a + bX$ and the proportion which is unexplained and is therefore due to the error variable E. (Again, we have considerable freedom in the way we choose to measure this variation. The most convenient way to define the total variation in the Y-values in Table 1 is to take it to be the sum of the squares of the differences between the Y-values and the average of the Y-values. Similarly, we can find the variation of the predicted Y-values. If we divide the second variation by the first one and multiply by 100, we have the percentage variation accounted for by the line of best fit.) In our example, the percentage variation explained by the line of best fit is 61.9% and the percentage unexplained by this relationship is 38.1%. Roughly speaking, this can be interpreted as saying that 61.9% of the first year Mathematics mark is decided by the 3-unit Mathematics mark and the remaining 38.1% by other factors. Thus we have obtained a fairly good fit, but it is not really good enough to give reliable predictions. Of course, we could now try to fit a more complicated function f(X) to the points in Figure 1, but this does not give any worthwhile improvement.

We can do a similar analysis using the Aggregate mark in the Higher School Certificate, instead of the 3-unit Mathematics mark. For the sample chosen, we found that only 36.1% of the first year Mathematics mark is accounted for by the Aggregate mark, so the Aggregate mark is even less useful as a predictor than the 3-unit mark. A word of caution, however, is appropriate. These analyses are based on a sample of only 20 students and therefore the conclusions should not be taken too seriously.

☆ ☆ ☆

MATHEMATICS (BE IN IT) COMPETITION

Competition number 2

There was a young fellow called Daedalus
Who claimed he'd invented the calculus.
But his work was derivative
And his integral primitive,
And it didn't make much sense at all to us.

For our second Mathematics (Be In It) Competition, you are asked to write a limerick with a mathematical theme. Prizes of \$3, \$2 and \$1 will be awarded for the best entries. Entries must reach the Editor by 1 October, 1979, and the prize-winners will be announced in Parabola, Volume 16, Number 1. Please make sure that your entry bears your name, address, year and school.

☆ ☆ ☆