DRAWING A STRAIGHT LINE THROUGH POINTS ON A GRAPH

By W. O. CHRISTIANSON

It is common practice for many people to feel that they can best illustrate the correlation between two sets of figures by plotting points on graph paper and then drawing a straight line through these points to express the average relationship between the two sets of data.

There is of course more than one way of drawing this line, the most popular being, apparently, by guessing where it should go. Unfortunately this guesswork can be most inaccurate and so the writer now presents a simple mathematical way of determining such a line with precision. The calculations involved may appear rather formidable but with a calculating machine they do not take very many minutes.

This mathematical process is called the method of least squares, the line so determined being such that the sum of the squared deviations between this line and the plotted points is at a minimum. The straight line so determined is called a 1st Order Polynomial Curve and the method of calculating this curve is conveniently illustrated by taking an example.

For such an example let us first of all examine the relationship between the Sucrose per cent Cane and the Sucrose per cent Bagasse figures shewn in the 34th Annual Summary of Chemical Laboratory Reports, Table V(1), from which the following data are taken.

	Sucrose per cent Cane	*Sucrose per cent Bagasse*	*Extraction*
May	11.60	2.16	93.33
June	12.48	2.34	93.33
July	13.26	2.53	93.25
August...	14.03	2.66	93.21
September	14.12	2.66	93.21
October	14.07	2.71	92.93
November	13.43	2.67	92.58
December	12.80	2.56	92.38
January	12.73	2.51	92.01
Total	118.52	22.80	836.23
Mean	13.1689	2.5333	92.9144

For our purpose we require to calculate for each set of the figures the total and mean (shewn above), sum of the squares of the deviations from the mean, and lastly the sum of the products of the corresponding deviations from the means.

Thus for Sucrose per cent Cane we calculate:

Total $=118.52$

Mean $=118.52 \div 9 = 13.1689$

Sum of squares $= \left\{(11.60)^2+(12.48)^2+(13.26)^2+ \ldots +(12.73)^2\right\} - \frac{(118.52)^2}{9}$

$=1566.5760-1560.7767$

$=5.7793$

Similarly for Sucrose per Cent Bagasse we have:

Total $=22.80$

Mean $=2.5333$

Sum of squares $= \left\{(2.16)^2+(2.34)^2+ \ldots +(2.51)^2\right\} - \frac{(22.80)^2}{9}$

$=0.2600$

Total sum of products of deviations from means

$$=(11.60\times2.16)+(12.48\times2.34)+ \ldots +(12.73\times2.51)-\frac{(118.52\times22.80)}{9}$$

$=+1.1434$

From the above we can now calculate a Regression Co-efficient and a Regression Formula to express the average relationship between our two sets of data.

Regression Co-efficient of Sucrose per cent Bagasse on Sucrose per cent Cane ... ... $= \frac{+1.1434}{5.7993}$

$= +0.1972$

The Regression Formula giving the average Sucrose per cent Bagasse for any value of Sucrose per cent Cane is calculated thus:

Sucrose per cent Bagasse ... = Mean Suc. % Bagasse + 0.1972 (Suc. % Cane − Mean Suc. % Cane)

Sucrose per cent Bagasse ... $=2.5333+0.1972$ (Suc. % Cane -13.1689)

$=2.5333-2.5969+0.1972$ (Suc. % Cane)

$= -0.0636 + .1973$ (Suc. % cane)

Hence for 11.00 Sucrose per cent Cane we calculate:

Sucrose per cent Bagasse ... $=-0.0636+0.1972$ (11.00)

$=2.11$

and for 14.00 Sucrose per cent Cane we calculate:

Sucrose per cent Bagasse ... $=-0.0636+0.1972$ (14.00)

$=2.70$

and so on.

We can now draw our straight line through points so calculated, as illustrated in the "graph" or more properly, the "scatter diagram", shewn in Figure 1.

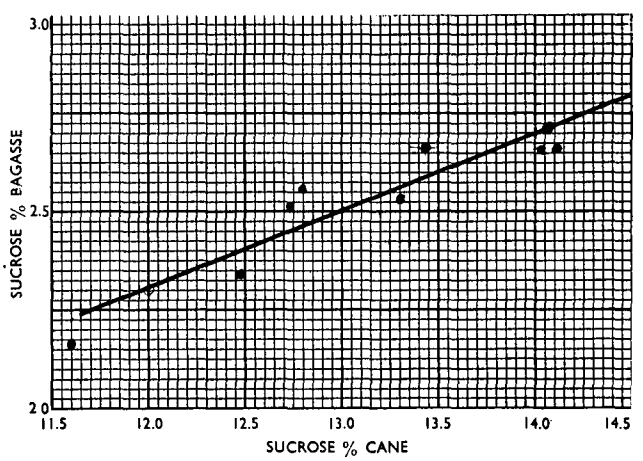

FIG. 1

But, and this is very important, has this line any meaning or in other words, does it reveal any real correlation between Sucrose per cent Cane and Sucrose per cent Bagasse? This question we can answer by calculating the Correlation Coefficient, (r), which is given by the formula:

$$r = \frac{\text{sum of products}}{\sqrt{\text{sum of sqs. of Suc. \% Cane} \times \text{sum of sqs. of Suc. \% Bagasse}}}$$

$$= \frac{+1.1434}{\sqrt{5.7993 \times 0.2600}} = +0.9312$$

By referring now to Prof. R. A. Fisher's table of r [2] we find that for 9 pairs of observations:

r is required to be at least 0.6664 at P=.05
and 0.7977 at P=.01

our Correlation Co-efficient of .9312 is therefore highly significant and we must therefore conclude that there is a definite association between Sucrose per cent Cane and Sucrose per cent Bagasse in the data we have dealt with.

A similar examination of the Sucrose per cent Cane and Extraction figures in Perk's table[1] leads to the Regression Equation:

Extraction = 92.9144 − 0.3371 + 0.0256 (Suc. % Cane).
= 92.5773 + 0.0256 (Suc. % Cane)

The line drawn from this formula is shewn in the scatter difigram illustrated in Figure 2.

The Correlation Co-efficient for the Sucrose per cent Cane and Extraction figures is however only 0.057, which from Fisher's table of "r" is not significant. We therefore must conclude that there is no significant association between Sucrose per cent Cane and Extraction in the data examined, even though the line slopes. (This slope has been accentuated, of course, in Figure 2 by suitably spacing the ordinates).

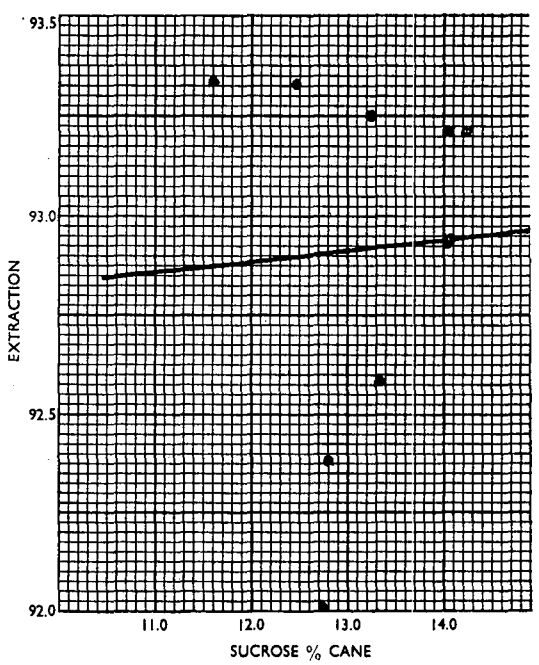

FIG. II

In the above the writer has taken all legitimate short cuts and presents a simple mechanical process as simply as possible. This has been done in the hope that the use of this mathematical method will find more general use, and also for simplicity, he has purposely left out a lot of detail and discussion which would only tend to confuse.

REFERENCES

(1) Perk, C. G. M., 34th Annual Summary of Chemical Laboratory Reports. Proc. S.A. Sugar Tech. Association, 33, 1959.

(2) R. A. Fisher: Statistical Methods for Research Workers.

Dr. Douwes-Dekker asked if the figures extracted from Perk's table were those of one factory or the averages from all factories.

Mr. Christianson replied that Perk's Table V showed the monthly averages for all factories in South Africa.

Mr. Beesley said that he was glad that Mr. Christianson had brought the method of least squares to notice, as it was applicable to many problems in a sugar factory, for instance, he had applied it to find molasses purity, and to Nutch purity vs. C massecuite purity, to determine the average C. massecuite purity that would give best molasses extraction for a given C massecuite station.

He said that while appreciating that Fig. II was meant mainly as an illustration of lack of correlation, he felt that the choice of variables (extraction vs. sucrose per cent cane) was most unfortunate, as he believed that sucrose per cent cane did have an effect on extraction. However in investigating the relationship it was necessary to study the effect of both sucrose per cent cane and fibre per cent cane at the same time, and preferable to follow the course of individual mills rather than the whole industry. He had developed a formula along these lines and it appeared to apply quite well to Tongaat, Natal Estates and Renishaw.

Mr. Christianson said that as sucrose per cent cane and sucrose per cent bagasse were so intimately associated one could not expect to obtain an increase in extraction with increase in sucrose per cent cane.

Mr. du Toit stated that the method outlined in the paper should be applied more generally than it was at present. He agreed with the results of the two correlations shown in the paper and said that Mr. Beesley's multiple correlation should be checked by calculating partial correlation coefficients. After eliminating the effect of fibre, he was convinced that no correlation would be found between sucrose per cent cane and extraction.

Mr. Rault enquired if it was accepted that each part of fibre would have the same effect on extraction, and if in the comparison of sucrose per cent cane with extraction, any other factors were taken into consideration.

Mr. Christianson replied that quality of fibre as well as the quantity of fibre must be taken into consideration but there was no doubt however that on the average the quantity of fibre was inversely associated with extraction. The comparison of sucrose with extraction shown in Figure 2 was a simple straightforward correlation and no other factors were taken into account.