attempted to write fairly large programmes can appreciate the difficulty in locating and removing errors without disturbing other functioning parts of the programme.

Programmers are also hindered by machine breakdowns and, especially in university computing departments, overstrained resources lead to too many demands on them.

### *Conclusion*

'Too often it is presumed that the amassing of large volumes of unselected data will allow its "digestion" by the computer, with the emergence of correlations, tabulations or hypothesis more or less at the touch of a button. Nothing could be further from the truth' (Taylor, 1967).

The use of a large computer in survey work does not at first appear to save a significant amount of time. Only when programmes are established and error-free and the collection and preparation of data are directed towards the method of analysis does the real time-saving become apparent.

Much time and thought is given to the feasibility and accuracy of sophisticated survey techniques. Unfortunately data which have been painstakingly collected may wait many months for analysis unless basic requirements of data-processing are anticipated and careful consideration given to relatively mundane aspects of survey planning.

The clearer the aims of a project and the closer the attention to the design of the data-recording sheet, the quicker will be the analysis and the smaller the output.

The password to successful computer-processing is, 'First find your programme'.

REFERENCES

Miller, D. S. (1965). *Progress in Medical Computing* p. 33. London: Elliott Medical Automation Ltd.

Nicholls, J. (1967). *Computer Weekly, London*, 5 October.

Rees, D. J. (1965). *Computer Unit Report* no. 2 (revised ed.). University of Edinburgh.

Taylor, T. R. (1967). *The Principles of Medical Computing*. Oxford and Edinburgh: Blackwell Scientific Publications.

## Statistics and computers—a worked example

By R. Thompson, *Agricultural Research Council Unit of Statistics, Edinburgh*

Computation plays a large part in applied statistical work and it is no wonder that the advent of high-speed computers has had an impact on its development. Yates (1966), indeed, believes that computers have started a second revolution in statistics. Computers have undoubtedly provided the impetus for research into new or previously unsolved problems in methodology but they can also provide speedier and more complete analysis using techniques known, if not used, in the desk calculator era.

I will illustrate various aspects of the latter use with data from two experiments (1 and 2) on the effect of stocking density on live-weight gain and feed conversion of broiler chickens. In Expt 1, eight pens of birds were reared from birth to slaughter at 10 weeks at each of four stocking rates (seventeen, twenty, twenty-three, twenty-six birds per pen). Unfortunately the birds could not be sexed at birth and so there were unequal numbers of male and female birds in each pen. Table 1 gives these numbers and also means of the logarithm of the 10-week weight, for healthy birds in each pen.

Table 1. *Pen means for healthy birds in Expt 1*

($Log_e$ (10-week weight in g) for females, males and also adjusted to a common proportion of males in each pen. The numbers of birds are in parentheses)

| Stocking rate 17 birds per pen | | | Stocking rate 20 birds per pen | | |
|---|---|---|---|---|---|
| Females | Males | Adjusted | Females | Males | Adjusted |
| 7·6212 (12) | 7·8344 (5) | 7·7159 (17) | 7·5886 (12) | 7·7501 (7) | 7·6637 (19) |
| 7·5717 (12) | 7·7657 (5) | 7·6608 (17) | 7·5402 (11) | 7·8342 (9) | 7·6700 (20) |
| 7·5718 (12) | 7·7526 (4) | 7·6587 (16) | 7·6179 (10) | 7·8150 (9) | 7·7035 (19) |
| 7·5616 (9) | 7·7851 (7) | 7·6597 (16) | 7·5542 (9) | 7·8534 (11) | 7·6941 (20) |
| 7·6321 (6) | 7·8391 (11) | 7·7199 (17) | 7·5967 (10) | 7·8390 (10) | 7·7043 (20) |
| 7·6594 (8) | 7·8192 (9) | 7·7239 (17) | 7·5905 (12) | 7·8592 (7) | 7·7051 (19) |
| 7·5670 (15) | 7·7288 (2) | 7·6571 (17) | 7·5296 (13) | 7·8860 (4) | 7·6585 (17) |
| 7·5983 (14) | 7·7981 (3) | 7·6916 (17) | 7·6437 (10) | 7·8556 (10) | 7·7361 (20) |
| Overall adjusted mean = 7·6864 ± 0·0098 | | | Overall adjusted mean = 7·6926 ± 0·0094 | | |
| Back transformed mean = 2175 g | | | Back transformed mean = 2189 g | | |

| Stocking rate 23 birds per pen | | | Stocking rate 26 birds per pen | | |
|---|---|---|---|---|---|
| Females | Males | Adjusted | Females | Males | Adjusted |
| 7·5966 (9) | 7·8137 (14) | 7·6912 (23) | 7·6543 (13) | 7·8082 (10) | 7·7221 (23) |
| 7·5734 (10) | 7·7708 (12) | 7·6574 (22) | 7·5928 (16) | 7·8013 (9) | 7·6853 (25) |
| 7·5565 (11) | 7·8220 (10) | 7·6746 (21) | 7·5751 (13) | 7·8221 (12) | 7·6845 (25) |
| 7·5197 (11) | 7·8311 (11) | 7·6618 (22) | 7·5640 (15) | 7·7950 (9) | 7·6647 (25) |
| 7·5717 (11) | 7·8469 (11) | 7·6957 (22) | 7·5942 (11) | 7·7473 (15) | 7·6519 (26) |
| 7·5779 (11) | 7·7756 (11) | 7·6632 (22) | 7·5696 (11) | 7·8276 (14) | 7·6872 (25) |
| 7·6379 (11) | 7·8150 (10) | 7·7139 (21) | 7·6613 (13) | 7·8117 (13) | 7·7230 (26) |
| 7·5934 (13) | 7·8561 (9) | 7·7073 (22) | 7·6215 (17) | 7·8133 (7) | 7·7100 (24) |
| Overall adjusted mean = 7·6831 ± 0·0091 | | | Overall adjusted mean = 7·6908 ± 0·0087 | | |
| Back transformed mean = 2168 g | | | Back transformed mean = 2185 g | | |

In Expt 2 again eight pens of birds were reared at four higher stocking rates (twenty-six, thirty-one, thirty-eight, fifty-two birds per pen). It was also decided to investigate the effect of two surfactants in this experiment. For this end, at each stocking density, four pens were fed a control diet and two pens were fed on the surfactant added diets. Table 2 gives the numbers of birds in Expt 2.

As a preliminary, before thorough analysis, checks were incorporated into the data to detect gross errors that might distort the analysis. For example, the 9- and 10-week weights for individual birds were compared, expecting the latter to be the larger.

Because editing procedures are tedious clerically, little work of this type used to be done, or only undertaken after analysis produced unexpected results. It is however

Table 2. *Numbers of female and male birds per pen in Expt 2*

| | Stocking rate 26 | | 31 | | 38 | | 52 | |
|---|---|---|---|---|---|---|---|---|
| | Female | Male | Female | Male | Female | Male | Female | Male |
| Diet 1 | 16 | 10 | 15 | 16 | 18 | 18 | 31 | 20 |
| | 11 | 13 | 18 | 12 | 21 | 15 | 30 | 21 |
| | 14 | 12 | 17 | 12 | 14 | 22 | 21 | 28 |
| | 13 | 13 | 14 | 15 | 18 | 18 | 25 | 25 |
| Diet 2 | 12 | 12 | 19 | 11 | 20 | 18 | 26 | 23 |
| | 15 | 11 | 14 | 17 | 20 | 16 | 22 | 26 |
| Diet 3 | 6 | 20 | 16 | 15 | 28 | 9 | 27 | 23 |
| | 14 | 10 | 17 | 13 | 21 | 16 | 27 | 24 |

simple to programme a computer to draw attention to anomalous data and if necessary take appropriate action.

Scatter diagrams often give clear indications of the underlying variation and enable outliers to be detected. Although equipment is now available for display of such diagrams on screens linked to computers, adequate results are often produced from line printer output. These are especially useful when many variables are considered together, for scatter diagrams often suggest which variables need sophisticated analysis, e.g. the analysis of variance, estimation of regression relationships, or multivariate methods.

In Expt 1 plotting 10-week weights showed that not only were the means different for males and females but that males varied more than females. At this stage we could either continue the analysis using separate variances for males and females or try to find a transformation that made the within-sex variances homogeneous. On a logarithmic scale the within-sexes variances were found to be more homogeneous and the analysis is illustrated using this transformation.

Transformations are generally used to make interpretation simpler by, for example, introducing a scale on which factors do not interact, or to simplify analysis, as with the heterogeneous within-sexes variances. With desk calculators it is usually impossible to investigate more than a single transformation chosen according to a well-defined rule of thumb. Computers make it easier to consider families of transformations and to investigate more thoroughly their precise effect on the resulting analysis (Box & Cox, 1964).

Graphical methods were also used to investigate whether within-pen variation depended on the number of birds per pen. For instance a peck order might arise in a heavily stocked pen with the bigger birds eating more food at the expense of the smaller birds. Plotting the within-pen variance against the number of birds in a pen did not show any evidence of such a relationship.

We now come to the main part of the analysis of Expt 1. We cannot compare stocking density means directly for differences between the means is due to variations from pen to pen in the proportions of males and females. So we first adjust each pen mean to the estimated value it would have had if the proportion of males had been equal to the overall proportion of males in the experiment.

From each pen we can estimate the difference due to sexes. Weighting each difference by (number of females × number of males)/number of birds, we can obtain the estimated sex difference for each stocking intensity and as these are not significantly different we use an overall sex difference for adjustment.

The adjusted pen means are given in Table 1. For example, for the first pen $7{\cdot}7159=(7{\cdot}6212\times 12+7{\cdot}8344\times 5)/17-0{\cdot}2214(5/17-290/661)$. Essentially when we make this adjustment we are saying that on a logarithmic scale a male is 0·2214 units heavier than a female, or on the untransformed scale a male is 1·247 times as heavy as a female.

The between-pen variation is needed for comparisons between stocking densities. The sum of squares for pens within the $j^{th}$ stocking density has expectation

$$\left(7+\left\{\sum_i^j (m_{ij}^2/p_{ij}) - (\sum_i^j m_{ij})^2/(\sum_i p_{ij})\right\}/w\right)\sigma_w^2 + \left(\sum_i^j p_{ij} - (\sum_i p_{ij})^2/(\sum_i^j p_{ij})\right)\sigma_p^2, \quad (1)$$

where $\sigma_w^2$ is the variance between birds, $\sigma_p^2$ is the variance between pens. $p_{ij}$, $m_{ij}$ are the number of birds and the number of male birds in the $i^{th}$ pen ($i=1, 2, \ldots 8$) on the $j^{th}$ stocking density and $W=\sum_i m_{ij}(p_{ij}-m_{ij})/p_{ij}$. The term involving W arises because we are adjusting to a constant proportion of males. It is in fact negligible in this example.

The variance of the $j^{th}$ stocking density mean is given by

$$\left\{1/(\sum_i^j p_{ij}) + \left[(\sum_i^j m_{ij}/\sum_i p_{ij}) - 290/661\right]^2/w\right\}\sigma_w^2 + \sum_i^j p_{ij}^2/(\sum_i^j p_{ij})^2\sigma_p^2.$$

The stocking density means and standard errors are given at the foot of Table 1 together with the back transformed means. Biologically, the latter can be more important.

In Expt 2 more complications arise owing to the introduction of the diet factor. Suppose we assume that the stocking density and the diets do not interact and that for any pen on stocking density i and diet j the adjusted pen mean $y_{ijk}$ can be expressed by the model

$$y_{ijk} = \mu + s_i + d_j + p_{ijk},$$

where $\mu$ is the general mean and $p_{ijk}$ is the residual variation, again made up of 2 parts due to the bird and pen variation. The estimation of the between-pen variation is slightly more difficult owing to the introduction of the diet factor.

The effects of stocking density and diets can be estimated from the least squares equation (2)

$$\begin{aligned}
1134\mu + 202s_1 + 241s_2 + 292s_3 + 399s_4 + 566d_1 + 282d_2 + 286d_3 &= 8595{\cdot}9159 \\
202\mu + 202s_1 \qquad\qquad\qquad\qquad + 102d_1 + 50d_2 + 50d_3 &= 1536{\cdot}5441 \\
241\mu \qquad + 241s_2 \qquad\qquad\qquad + 119d_1 + 61d_2 + 61d_3 &= 1829{\cdot}3710 \\
292\mu \qquad\qquad + 292s_3 \qquad\quad + 144d_1 + 74d_2 + 74d_3 &= 2213{\cdot}2652 \\
399\mu \qquad\qquad\qquad + 399s_4 + 201d_1 + 97d_2 + 101d_3 &= 3016{\cdot}7357 \\
566\mu + 102s_1 + 119s_2 + 114s_3 + 201s_4 + 566d_1 \qquad\qquad &= 4291{\cdot}8664 \\
282\mu + 50s_1 + 61s_2 + 74s_3 + 97s_4 \qquad + 282d_2 \qquad &= 2141{\cdot}0201 \\
286\mu + 50s_1 + 61s_2 + 74s_3 + 101s_4 \qquad\qquad + 286d_3 &= 2163{\cdot}0294
\end{aligned} \quad (2)$$

The computer can help not only in providing solutions to these equations, but it can also be programmed to form automatically the initial equations.

Although there are eight equations in (2) only six constants can be estimated since

https://doi.org/10.1079/PNS19700032 Published online by Cambridge University Press

the sum of the second, third, fourth and fifth and the sum of the last three equations give the first equation. We can put $\mu=0$ and $d_1=0$ in order to estimate the other constants. The choice of constraints is arbitrary and the sum of squares due to stocking density and diets is independent of the constraint introduced. However, the interpretation of the constants can change and with $d_1=0$ $d_2$ now estimates the difference between diet 1 and diet 2.

The precision of these estimates needs the calculation of the inverse of the matrix given by the coefficients of the parameters in equation (2). The inverse is set out in Table 3. Basically any row of the inverse is the solution of the set of equations (2), with a system of 0's and 1's instead of the observations on the right-hand side. The diagonal elements of Table 4 give the number of birds in a pen, squared for individual stocking densities and diets. The off-diagonal elements correspond to the similar sums for stocking density diet combinations.

Table 3. *Inverse arising from equations* (2) ($\times 10^6$)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 5814 | 884 | 883 | 867 | 0 | −1752 | −1740 |
| 0 | 884 | 5053 | 905 | 886 | 0 | −1792 | −1779 |
| 0 | 885 | 905 | 4331 | 887 | 0 | −1794 | −1782 |
| 0 | 867 | 886 | 887 | 3375 | 0 | −1739 | −1762 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | −1752 | −1792 | −1794 | −1739 | 0 | 5313 | 1767 |
| 0 | −1740 | −1779 | −1782 | −1762 | 0 | 1767 | 5263 |

Table 4. *Pen numbers, squared, summed over stocking density and diet combinations*

| | | | Density | | | | Diet | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Overall | 26 | 31 | 38 | 52 | 1 | 2 | 3 |
| Overall | | 42944 | 5108 | 7265 | 10662 | 19909 | 21434 | 10558 | 10952 |
| | 26 | 5108 | 5108 | 0 | 0 | 0 | 2604 | 1252 | 1252 |
| Density | 31 | 7265 | 0 | 7265 | 0 | 0 | 3543 | 1861 | 1861 |
| | 38 | 10662 | 0 | 0 | 10662 | 0 | 5184 | 2740 | 2738 |
| | 52 | 19909 | 0 | 0 | 0 | 19909 | 10103 | 4705 | 5101 |
| Diet | 1 | 21434 | 2604 | 3543 | 5184 | 10103 | 21434 | 0 | 0 |
| | 2 | 10558 | 1252 | 1861 | 2740 | 4705 | 0 | 10558 | 0 |
| | 3 | 10952 | 1252 | 1861 | 2738 | 5101 | 0 | 0 | 10952 |

The pen variation can be estimated by equating the difference between the pen sum of squares and the diet-stocking density sum of squares and its expectation. If $A_{ij}$ and $B_{ij}$ are the elements on the $i^{th}$ row and $j^{th}$ column of Table 3 and Table 4 the coefficient of pen variance in the equation similar to (1) is then given by

$$\text{total number of observations} - {}_i, \Sigma_j A_{ij} B_{ij}.$$

Similar results can be derived if we introduce an interaction term into the model and this was, in fact, used in a complete analysis.

The theory in the preceding paragraphs has been given in a notation which hand calculators can comprehend. For example, the weighting factor approach used in Expt 1 is just algebraic manipulation of least squares theory into a form suitable for hand computation. It would be laborious, but not impossible, to extend this approach to adjust for three or more factors. However, least squares theory can be expressed more concisely in matrix algebra. This type of algebra can easily be translated into terms the electronic computer can understand. So it becomes relatively easy to adjust for more factors or analyse additional variates, such as 8- and 9-week weight.

An interesting development is described by Hartley (1967). He was concerned with finding the coefficients of the variance components such as in formula (1) earlier. Even in matrix algebra the formulae for these coefficients can be very complicated for uncommon designs. For large experiments, due to the size of matrices used, they might have needed to be computed piecemeal and need special programming skill.

Hartley shows that the coefficients can be expressed in terms of the sums of squares from an analysis of variance carried out on certain dummy variates. The beauty of this approach is that we only need the computer programme for the analysis of variance. For Expt 1 it would require thirty-two analyses of variance to obtain the pen coefficient. This approach is specifically for the electronic computer, not for the desk calculator.

I would emphasize that the fact that computers can solve large sets of equations with ease should be used with discrimination and does not give *carte blanche* for the use of unbalanced designs. Essentially the two experiments were balanced with respect to the treatment structure. The non-orthogonality arose only because of its being unpractical to sex birds at birth, and because odd birds died. Usually experimenters are interested in making a certain number of treatment comparisons with equal accuracy. In this case, balanced designs usually are suggested, because they fulfil this requirement, and not, although it usually happens, to simplify the resultant calculations. Interpretation is more difficult from non-orthogonal experiments. I would rather look on general regression techniques as salvage tools when accidents occur, i.e. animals die, plots fail, than as sensitive instruments for unthinking experimenters.

In this paper I have been concerned with the analysis of data rather than with design of experiments but there are possibilities in this field. For example, when good experimental practice suggests grouping together in blocks, say milking cows of similar live-weight, there has often been a compromise to be made between using all animals in blocks of unequal sizes and ignoring a few animals and using more symmetrical designs. The choice of design in these cases seems an ideal task for the computer to tackle because of the large number of combinations needing consideration with any moderately unbalanced experiment. Once such algorithms are available it seems logical that the next step would be to link these to a programme that produces an acceptable randomization of the design. Owing to the complementary nature of the design and analysis of experiments, it will often happen that any programme used to investigate the properties of a set of designs will, with little modification, actually analyse the resultant data. This can often save time, especially if an intricate sequence of experiments is in progress and one requires the analysis of one set of experiments before deciding on the exact strategy of the next set.

I would like to thank Dr Bolton and Dr Dewar of the Poultry Research Centre for permission to use their data.

REFERENCES

Box, G. E. P. & Cox, D. R. (1964). *Jl R. statist. Soc.* B, **26**, 211.
Hartley, H. O. (1967). *Biometrics* **23**, 105.
Yates, F. (1966). *Biometrics* **23**, 233.