Lih, H. & Baumann, C. A. (1951). *J. Nutr.* **45**, 143.
Sauberlich, H. E. (1952). *J. Nutr.* **46**, 99.
Schendel, H. E. & Johnson, B. C. (1954*a*). *Fed. Proc.* **13**, 623.
Schendel, H. E. & Johnson, B. C. (1954*b*). *J. Nutr.* **54**, 461.
Waibel, P. E., Bird, H. R. & Baumann, C. A. (1954). *J. Nutr.* **52**, 273.

The use of an electronic computer in the estimation of sampling errors in a nutritional survey

By B. M. CHURCH and S. LIPTON

Rothamsted Experimental Station, Harpenden, Herts

(*Received* 3 *October* 1955)

If observations covering the whole of a body of material are available, the errors which are to be expected when sampling further material of the same type can be estimated and the relative accuracy of different sampling methods determined. Unfortunately, however, the numerical work required for investigations of this kind is very considerable, and in consequence when only desk machines are available statisticians are often reluctant to undertake this type of investigation with sufficient thoroughness to build up an adequate body of knowledge on sampling errors.

The advent of electronic computers has completely changed the situation, for it is now possible to carry out the required numerical calculations without difficulty. Since work of this kind is essentially repetitive it is eminently suitable for electronic computers, for once a set of instructions (technically known as a programme) appropriate to a particular problem has been written for a machine, further data of the same type can be analysed with little effort.

The present paper deals with the relatively simple but important problem of sampling numerical sequences. The investigation was made in order to determine the number and length of the periods over which an individual's food intake should be observed in order to give sufficiently accurate assessment of the intake of the various dietary components. The nutritional aspects of the investigation and the results are reported by Chappell (1955). Since sampling problems of this type frequently arise in nutritional work and in other fields, it is considered that a brief description of the procedure adopted will be of value. Now that electronic methods of computation are available this type of calculation should become a standard routine when thorough knowledge of sampling errors is required for planning further work.

The problem

In sampling a numerical sequence we can vary both the size of the sampling unit (here defined as the length of sequence, or number of terms r, included in each unit) and the number of such units sampled. Thus in the nutritional problem periods of, for

example, a week, 2 weeks, a month may be taken as sampling units, and varying numbers of these units selected.

Three methods of selecting the required units are of interest:

(*a*) Two units may be selected at random from 'blocks' (or strata) of the appropriate size.

(*b*) One unit may be selected at random from blocks of half the size of those required for method (*a*). (Blocks of this size are here defined as containing t terms or t/r sampling units.)

(*c*) Units may be selected at regular intervals of t terms or t/r sampling units.

Methods (*a*) and (*b*) are known as stratified random sampling with two and one units per block respectively. Method (*c*) is known as systematic sampling. Method (*a*) has the advantage that sampling errors can be determined from the data of the sample only. Method (*b*) also has a well-defined sampling error which will almost always be less than that of method (*a*), but this cannot be simply or accurately determined from the data provided by a particular sample. In method (*c*) the sampling errors are uncertain since the necessary random components of selection are lacking. In particular, the errors will clearly be large if any periodicities or quasi-periodicities of the same period as the sampling interval are present. Nevertheless, systematic samples are frequently used because of their practical convenience and, if no periodicities are present, they are likely to be somewhat more accurate than the samples of method (*b*) (Yates, 1948). In general, it will be adequate to take the sampling errors of method (*c*) as equivalent to those of method (*b*).

To evaluate the sampling error of method (*b*) for sampling units of 1 week and blocks of 8 weeks given, say, data for 64 weeks, the data may be divided into 8 blocks of 8 weeks; the sum of the squares of deviations of the individual weekly values from the block means can then be calculated. Although this calculation is not difficult, the labour rapidly mounts up when different block sizes and different unit sizes require investigation.

Moreover, there is a further difficulty, concealed in the above example but very evident if, say, data for only 63 weeks are available: there will then be room for only 7 blocks of 8 and the exact location of these blocks is clearly arbitrary. In general, for a sequence with arbitrary starting and end-points, no particular location is to be preferred and logically all possible locations have an equal right to inclusion, though the additional information thereby obtained on the sampling errors is of course by no means proportional to the number of locations included.

These difficulties can all be resolved, however, by taking the smallest sampling units which require consideration (here 1 week), evaluating the differences of units one apart, two apart, and so on, and calculating the mean squares of these differences. The sampling errors appropriate to different sizes of unit and sizes of block can then be calculated from the formulas given in the next section. This method has incidentally the important additional advantage that proper account can be taken of missing values without undue labour.

Variance estimates

The variance between the terms, x_i, of a sequence within a group of t consecutive terms is given by

$$\frac{1}{t-1}\left\{x_1^2+\ldots+x_t^2-\frac{1}{t}(x_1+\ldots+x_t)^2\right\}, \quad (1)$$

which may alternatively be written

$$\frac{1}{t(t-1)}\{(x_1-x_2)^2+(x_1-x_3)^2+\ldots+(x_2-x_3)^2+\ldots+(x_{t-1}-x_t)^2\}. \quad (2)$$

The average variance within all such groups of t terms can therefore be estimated as

$$\frac{1}{t(t-1)}\{(t-1)\,d_1^2+(t-2)\,d_2^2+\ldots+d_{t-1}^2\}, \quad (3)$$

where d_i^2 is the mean value of the square of the difference between values i apart in the sequence, given by the equation

$$d_i^2=\frac{1}{t-i}\{(x_1-x_{1+i})^2+(x_2-x_{2+i})^2+(x_3-x_{3+i})^2+\ldots+(x_{t-i}-x_t)^2\}.$$

Variance estimates derived from equation (2) using all possible groups of t consecutive terms, and those from equation (3), are not identical for finite sequences since terms at the end of the sequence occur in a smaller number of groups. However, equation (3) is equally appropriate if the observed sequence is an arbitrary section of a longer sequence, and if there is no reason to expect different variability at the ends of the observed sequence. Under these conditions, it follows from equation (3) that the variance between groups of r consecutive terms within groups of t terms ${}_t\sigma_r^2$ (on a per term basis) can be estimated as

$${}_ts_r^2=\frac{1}{(t-r)}\left\{\frac{r}{t}[(t-1)\,d_1^2+(t-2)\,d_2^2+\ldots+d_{t-1}^2]-\frac{t}{r}[(r-1)\,d_1^2+(r-2)\,d_2^2+\ldots+d_{r-1}^2]\right\}. \quad (4)$$

The behaviour of equation (4) when these conditions are not satisfied has not been considered and might merit further investigation; the equation appears to be satisfactory for the present data and has been used in the programme below. If variance estimates were calculated from expressions in the form of equation (2) the arithmetic would be complicated if occasional terms were missing from the sequence; however, working from equation (4), it is only necessary to note the numbers of differences which are known and to use these as divisors when calculating d_i^2.

For long sequences, use of equation (4) is equivalent to estimating the variance between groups of r terms within all possible overlapping groups of t terms and, when the x_i are independently and normally distributed, provides $3r^2/(2r^2+1)$ times as much information as the estimate from a single set of non-overlapping groups of t terms.

Method of calculation

The electronic computer used for these calculations was the '401' prototype model built by Elliott Brothers under contract from the National Research Development Corporation (Lipton, 1955). This computer operates with numbers of 32 binary digits

(one of which is a sign digit), corresponding to about $9\frac{1}{2}$ decimal digits. The information is fed into the machine by means of five-hole punched tape and the output is by electric typewriter or teleprinter punch. As the various types of electronic computers have different order codes (i.e. methods of instructing the machine what to do) as well as different characteristics, a detailed description of the programme would only have a very limited interest. Consequently, an outline of the general method of work is given below, and details of actual coding and of points peculiar to the '401' are omitted.

The data from the dietary survey consisted of sixty-one observations on weekly intakes of each of thirteen dietary constituents relating to a total period of 63 weeks—in every instance the records for the 10th and 40th weeks were missing.

The data for each of the constituents were punched on a separate tape; every constituent had a code number and it, too, was punched on the tape and preceded the data proper. The code number was printed out by the machine at the head of the corresponding results and so prevented the possibility of identifying a set of results with the incorrect constituent. The punching of the data tapes was straightforward, the numbers being punched successively in order from the 1st week up to the 63rd week. When a week for which the observation was missing was reached, a large number (10^7) was punched—it was much greater than any actual observation and served as a signal to the machine that there was no record for that week.

The programme consisted of two stages. The machine was instructed to calculate and print out

(1) the d_i^2 for terms 1, 2, ..., $m-1$ apart;

(2) the sums $T_i=(i-1)\,d_1^2+(i-2)\,d_2^2+\ldots+d_{i-1}^2$ for $i=2, 3, \ldots, m$.

It is worth noting that although the T_i's could be computed on a desk machine by using the values of the d_i^2, this work would be time-consuming as well as boring. This illustrates an important advantage of electronic computers; as far as is required, subsequent minor calculations can be carried out by the machine in addition to the main calculations so that the results are obtained in the most convenient form.

The programme was written so that it could be used for any similar problem with an arbitrary number of observations limited only by the storage capacity of the machine. When the general programme tape is prepared, all that is required are the data tapes and a small tape for setting n, the total number of observations, and $m-1$, the maximum distance apart of the terms used in the d_i^2. There is no need to set parameters to indicate the number or positions of the missing values: as the machine forms each difference (x_j-x_{j+i}), x_j and x_{j+i} are tested to see if they represent genuine observations or missing values. If either represents a missing value, the machine passes on to the next difference and the divisor to be used in calculating d_i^2 is reduced by one.

In the dietary problem $n=63$ and $m=52$; thus 102 results had to be printed out for each variate. Subsequently ${}_t\sigma_r^2$ was estimated as

$$ {}_ts_r^2=\frac{1}{t-r}\left\{\frac{r}{t}\,T_t-\frac{t}{r}\,T_r\right\},$$

which was computed on a desk calculator.

Computing time

The entire procedure from the time the data tape was fed in until the printing of the final value for $i=52$ was carried out without the intervention of the operator. The machine took 4 min for each tape, composed as follows:

reading in data tape	10 sec
computing	$2\frac{1}{2}$ min
printing out results	$1\frac{1}{2}$ min.

Punching of each data tape, including duplication and checking, took about 10 min. In the general instance the computing time would be roughly proportional to $m(2n-m-1)$.

The whole programme, including routines for taking in the data and printing the results, occupied some 300 locations of the store out of a total of nearly 3000 in the computer; thus there was ample room for further programmes that could operate on the results.

Example

An extract from the figures printed by the electronic computer for one of the thirteen dietary constituents (total calories) is given in the second and third columns of Table 1. At the right of the table the variance estimates subsequently calculated are shown. The use of such variance estimates may be illustrated as follows. If the

Table 1. *Sampling variance of calorie intake (on a per week basis) calculated for various sizes of unit and of strata*

From electronic computer			Subsequent calculation of variance per week. Size of strata (t weeks)				
i	d_i^2	T_i	Size of sample unit (r weeks)	4	8	13	26	52
				Variance per week $\frac{1}{1-r}\left\{\frac{r}{t}T_t-\frac{t}{r}T_r\right\}$				
1	41,741		1	24,143	25,061	25,405	27,300	29,608
2	56,530	41,741	2	30,688	30,649	30,764	34,266	38,695
3	51,434	140,012	4	—	30,571	30,873	38,494	47,369
4	47,757	289,717	6	—	—	32,769	44,775	57,846
5	53,715	487,179	8	—	—	—	49,939	67,220
6	55,246	738,356						
7	52,235	1,044,779						
8	48,840	1,403,437						
⋮								
13	60,625	3,963,173						
⋮								
26	68,875	17,745,215						
⋮								
52	—	78,520,158						

average weekly calorie intake over a period of 2 months is to be determined, a single random sample of 1 fortnight would provide an estimate with standard error ± 124 ($=\sqrt{[30{,}649/2]}$), whereas 2 weeks sampled independently at random within consecutive months would give an estimate with standard error ± 110 ($=\sqrt{[24{,}143/2]}$), providing $\frac{30{,}649}{24{,}143}=1{\cdot}27$ times as much information as the single sample of the same total duration.

SUMMARY

1. The empirical evaluation of sampling errors of all kinds has in the past been greatly neglected owing to the laborious calculations required on desk machines. With the development of electronic methods these computations can now be readily undertaken.

2. The paper describes a method of evaluating the sampling errors of a numerical sequence using the electronic computer installed at Rothamsted Experimental Station. The method was developed to deal with the data of a dietary survey, but the programme has been written in general form and can take account of missing values.

The authors are indebted to Dr F. Yates, F.R.S., whose interest stimulated the investigation, for advice in the preparation of this paper.

REFERENCES

Chappell, G. M. (1955). *Brit. J. Nutr.* **9**, 323.
Lipton, S. (1955). *Mathematical Tables and Other Aids to Computation*, **9**, 69.
Yates, F. (1948). *Phil. Trans.* A, **241**, 345.

The role of fat in the diet of rats

8. Influence on growth of shortening products, 'emulsifier PT 006' and polymerized linseed oil

BY E. AAES-JØRGENSEN, J. P. FUNCH, P. F. ENGEL AND H. DAM

Department of Biochemistry and Nutrition, Polytechnic Institute, Copenhagen

(*Received* 5 *October* 1955)

In a previous study (Aaes-Jørgensen, 1954) polymerized herring oil given as the sole fat in the diet of newly weaned male rats was found to be toxic. At the level of 7% it depressed growth. At a 28% level the animals were dying after 14 weeks of feeding. Diarrhoea was not a striking sign. In a subsequent experiment a partial polyglycerol ester of in vacuo polymerized soya-bean oil, PT 006 (1 and 3·5% of the diets) was mixed with lard and fed to newly weaned rats in synthetic diets containing 28% lard plus emulsifier. The growth of these animals was almost the same as that of the controls given 28% lard throughout an experimental period of 15 weeks. At the end of this period the animals were killed. At autopsy no signs of carcinomatous tissue were found in the digestive tract.

The present experiments were carried out to study the effect on young rats of diets containing polymerized oils, in particular the effect of a polymerized oil used as an emulsifier in a shortening*, and of the emulsifier itself as the predominant dietary fat component.

* Shortenings (or compounds) are fats used to make pastry, cakes and such-like short, i.e. breaking or crumbling readily.