The Annals of Human Genetics has an archive of material originally published in print format by the Annals of Eugenics (1925-1954). This material is available in specialised libraries and archives. We believe there is a clear academic interest in making this historical material more widely available to a scholarly audience online.
These articles have been made available online, by the Annals of Human Genetics, UCL and Blackwell Publishing Ltd strictly for historical and academic reasons. The work of eugenicists was often pervaded by prejudice against racial, ethnic and disabled groups. Publication of this material online is for scholarly research purposes is not an endorsement or promotion of the views expressed in any of these articles or eugenics in general. All articles are published in full, except where necessary to protect individual privacy.
We welcome your comments about this archive and its online publication.

SOME EXAMPLES OF BIASED SAMPLING

By F. YATES

I. General

Sampling as ordinarily carried out is very frequently biased. Kiser(2) has recently given an account of a case where considerable sampling biases arose in a population study owing to the failure of enumerators to revisit missed families. The purpose of the present note is to illustrate the very large biases that can arise in the sampling of agricultural (and other) material if a proper process of selection is not laid down and adhered to.

With the spread of biological experimentation and observation sampling technique has assumed considerable importance, both as a method of avoiding the frequently excessive task of measuring or observing the whole of the available material and also as a method of securing data which involves the destruction of the observed material without seriously disturbing the course of the experiment. Sampling is also widely used in the study of economic and sociological problems, in the control of quality in modern mass production, and as a quality indicator in the buying of goods in bulk.

The ideal which is aimed at in sampling is to make the sample as representative as possible, so that measurements or observations on it can be taken as virtually equivalent to similar measurements on the whole population. The fact that this ideal is in the mind of the sampler when taking the sample naturally influences his selection if he has any freedom of choice. Most samplers when selecting a representative sample will deliberately reject abnormal material, or if they feel that the sample should be representative of the abnormal as well as the normal will deliberately balance up the different categories of abnormality.

Unfortunately the sampler's claims to be able to select "a representative sample" by personal judgment are largely unfounded, and his selection is in fact subject to all sorts of biases, psychological and physical. To avoid these biases and to provide an estimate of the representativeness of the sample, i.e. of the "sampling error", more rigorous processes of selection have been devised.

The following conditions must be satisfied:

(1) If bias is to be avoided, the selection of the samples must be determined by some process uninfluenced by the qualities of the objects sampled and free from any element of choice on the part of the observer.

(2) If a valid estimate of sampling error is to be available, each batch of material must be so sampled that two or more sampling units are obtained from it. These sampling units must be a random selection from the whole aggregate of sampling units that can be taken from the batch of material, and all the sampling units in the aggregate must be of approximately the same size and pattern, and must together comprise the whole of the batch of material.

The determination of the sampling error is important in order to ascertain the amount of sampling that is necessary to give the required accuracy. It is often thought that the ideal is to sample the whole population and that every effort should be made to approach this ideal as closely as possible. In reality, however, other sources of variation usually affect the results (or their applications), and there is little to be gained by reducing the sampling error to less than 25 per cent. of the other errors. The total error including sampling will then be approximately $\sqrt{(100^2 + 25^2)}$, or 103·1 per cent. of the error, if the whole population were sampled, and the loss of information (which is inversely proportional to the square of the standard error) is only 6 per cent. It is vital to keep this consideration in mind if the sampling is to be carried out both expeditiously and efficiently.

If condition (2) is fulfilled, condition (1) is automatically fulfilled also. There is nothing in the second condition which prevents the sampling unit being of a complex nature. Thus in a house to house survey with streets of approximately equal length the sampling unit might consist of every 10th house in a street. The aggregate of sampling units would then be houses 1, 11, 21, ..., houses 2, 12, 22, ..., etc., and the random selection of two numbers between 1 and 10 would effect the selection of two samples from the aggregate of the ten samples. The sampling error would be estimated from the difference of the means of the two sampling units. Each street would provide one such difference, and the final estimate would be made by combining all these. This can most conveniently be done by the methods of the analysis of variance.

There is also nothing in the second condition that prevents the batches of material which are sampled being subdivisions of the groups in which we are interested. Thus in a house to house survey we might arrange to sample by streets (every street being sampled), although our interest lay in districts. Equally we might subdivide the streets and sample within the subdivisions.

The elucidation of the various aspects of the sampling problem and of the conditions necessary for a valid estimate of sampling error is primarily due to R. A. Fisher; an adequate sampling technique for the sampling of growing crops was first developed by him in conjunction with other workers at Rothamsted. A brief account of this technique is given by Fisher himself in the discussion on the paper by Neyman referred to below. The principles there described are equally applicable to other types of sampling, as Fisher points out.

Randomisation is sometimes unfairly criticised for giving inaccurate samples. Streets, for example, may vary considerably from end to end, and if this is the case the selection of a number of houses at random from a whole street will undoubtedly give a more variable sample than would the selection of houses at equal intervals along the street. The remedy, however, lies not in abandoning the random process but in imposing restrictions (such as those suggested above) which do not conflict with condition (2). Some element of randomisation, therefore, is essential if there is to be a valid estimate of error, and is by far the simplest way of ensuring freedom from bias. But it must be properly performed by the use of a table of random numbers, well-shuffled cards, or similar apparatus. To select samples

"at random", or by the adoption of such processes as casting a hoop (or equally to think of numbers and call them "random numbers"), is emphatically not good enough, and may give rise to considerable biases, as will be seen from the examples given below.

A rather different type of selection, which does not depend so much on the choice of the observer, is that of making the means of one or more easily determinable characteristics of the sample agree as nearly as possible with the corresponding means of the population; this process has been called "purposive selection". Thus in sampling all the families of a district the sample might be so arranged that the mean number of children per family was the same as that of all the families in the district (as determined by previous census returns).

This method of sampling has recently been compared with the method of random sampling (including random sampling of a complex nature) in an excellent paper by Neyman (3) read before the Royal Statistical Society. Neyman comes to the conclusion that the method of purposive selection is definitely unreliable. I cannot do better than quote his own words:

"On the other hand, the consistency of the estimate... based upon a purposely selected sample depends upon hypotheses which it is impossible to test except by an extensive enquiry.

"If these hypotheses are not satisfied, which I think is a rather general case, we are not able to appreciate the accuracy of the results obtained. Thus this is not what I should call a representative method. Of course it may give sometimes perfect results, but these will be due rather to the uncontrollable intuition of the investigator and good luck than to the method itself. Even if the underlying hypotheses are satisfied, we have to remember that the elements of sampling, which it is possible to use when applying the purposive selective method, must be very few in number and very large in size. Consequently I think that when using this method we are very much in a position of a gambler, betting at one time £100."

Frequently the arrangement of a process of selection which is really random requires considerable ingenuity. Sometimes it is impossible or completely impractical, but in such cases all samples must be regarded with a greater or less degree of suspicion, and if possible regular measures of their reliability should be arranged by the occasional complete measurement of a whole batch of data.

II. Biases in height measurements on wheat

Our first two examples occurred in the sampling observations on the growth of the wheat plant which are now in progress at ten centres in the United Kingdom under the Ministry of Agriculture's crop-weather scheme.

The observations are carried out as follows. Two or three varieties are grown in eight blocks, one plot of each variety in each block, the varieties being arranged at random within the blocks. Each plot is 25 metres long and one drill width wide; it is normally divided into two half-drill widths of six rows each, one row being omitted down the centre of the plot. A sampling unit consists of four contiguous quarter-metre lengths of drill row taken from

the four central rows of the half-drill width. Height measurements are made on the shoot at each end of each sample quarter metre, eight shoots in all per sample. Two samples are taken from each half-drill width at each time of observation. To locate the samples two numbers between 1 and 100 are taken at random by means of a table of random numbers, and these are taken as the serial numbers of the two quarter metres (counting from one end of the plot) that are to be sampled. The actual location is by pacing. A U-shaped rod with its arms 25 cm. apart is used to demarcate the whole of the sample simultaneously, being thrust across the rows low down at right angles to them.

It will be seen that the requirements of condition (2) are rigorously satisfied by this sampling process, and that any variation affecting the whole of a half-plot is eliminated from sampling error.

At one of the centres only three rows were available for sampling (owing to the use of a small seed drill), and there were therefore only six shoots available for height measurements in each sample. In order to obtain a full eight measurements the observer selected the two additional shoots "at random" by eye. Fortunately the observations were booked in such a manner that it was possible to distinguish these additional measurements.

Table I. *Height measurements of wheat. Distribution of additional observations relative to the regular observations*

Date	Observation	Ascending order of magnitude								Total*	Expectation in each class
		1	2	3	4	5	6	7	8		
May 31	Shoot height	9	7	11	8	11	18	21	31	116	$14\frac{1}{2}$
June 14	Shoot height	11	13	13	10	21	23	18	17	126	$15\frac{3}{4}$
June 28	Ear height	9	19	27	23	15	10	5	4	112	14

* Owing to the existence of blank rows in certain samples the full totals of 128 are not attained.

Since there are eight height measurements in a sample, including the two additional ones, the additional measurements may occupy any two positions from 1 to 8 when all the measurements are arranged in order of magnitude. If the additional observations were really random, then the positions 1–8 should occur with equal frequency. Table I shows the observed frequencies for shoot height on May 31 and June 14, and for ear height on June 28, these being the three sets of observations which were readily classifiable in this manner. Fig. 1 shows the same frequencies graphically. The deviations from equal frequency are striking. Equally striking are the differences in type of deviation at the different dates. On May 31, when the wheat was under 2 ft. high, there is apparently a simple tendency to select the taller shoots, with perhaps a special preference for the very tall shoots. On June 14, this tendency, though much less, is till noticeable, but there is now a tendency to avoid very high values. On June 28, when the wheat was about 4 ft. high, there is a marked avoidance of the tallest shoots, and also to a less extent of the shortest shoots. The general bias is now strongly negative.

The actual deviations of the six regular measurements of each sample from their own mean, and of the two additional measurements from the mean of the regular measurements of the sample, were calculated for the observations of May 31 and June 28. The distributions of these are shown in Figs. 2 and 3. In each case the shaded histogram represents the distribution of

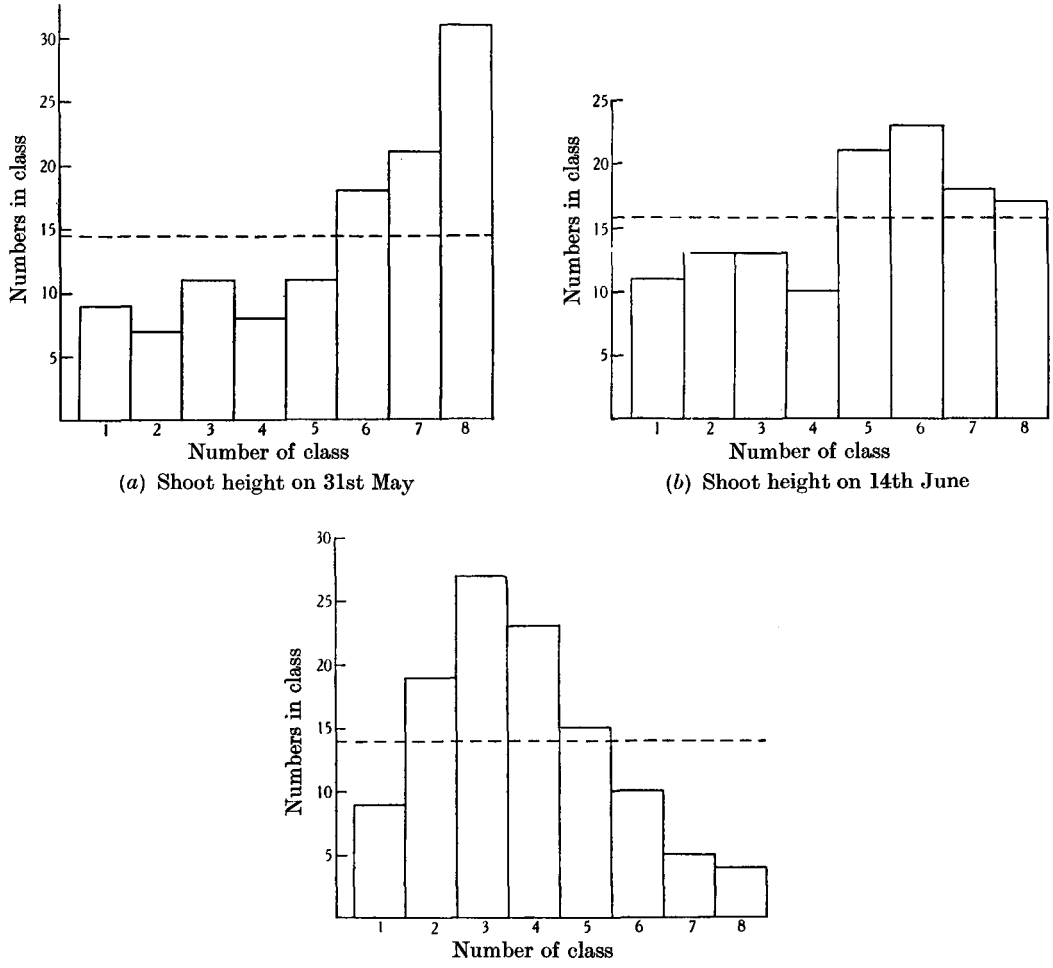

Fig. 1. Distribution in ascending order of magnitude of additional observations relative to the regular observations. The classes 1–8 are shown on the horizontal scales, and the numbers in each class on the vertical scales. The dotted lines show the expected values in each class on the hypothesis that the additional observations are truly random.

the regular measurements and the unshaded histogram that of the additional measurements. Since in a normal distribution with variance σ^2 the mean square of the deviations of six quantities from their own mean is $\frac{5}{6}\sigma^2$ and of the deviation of any other observation from the mean of these six is $\frac{7}{6}\sigma^2$, the horizontal scale of the additional observations has been reduced in the ratio $\sqrt{5}:\sqrt{7}$, so that the diagrams represent fairly the bias in variability in addition

to the bias in the mean. The number of additional observations falling into each class has been multiplied by three so as to give each histogram the same total area. The positive bias on May 31, and the negative bias and avoidance of extreme values on June 28, are both apparent in the diagrams.

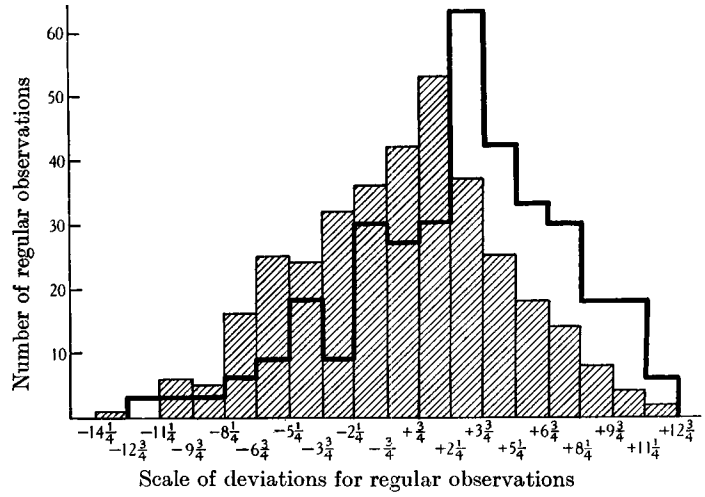

Fig. 2. Distribution of regular observations (shaded) and additional observations (unshaded) on 31st May.

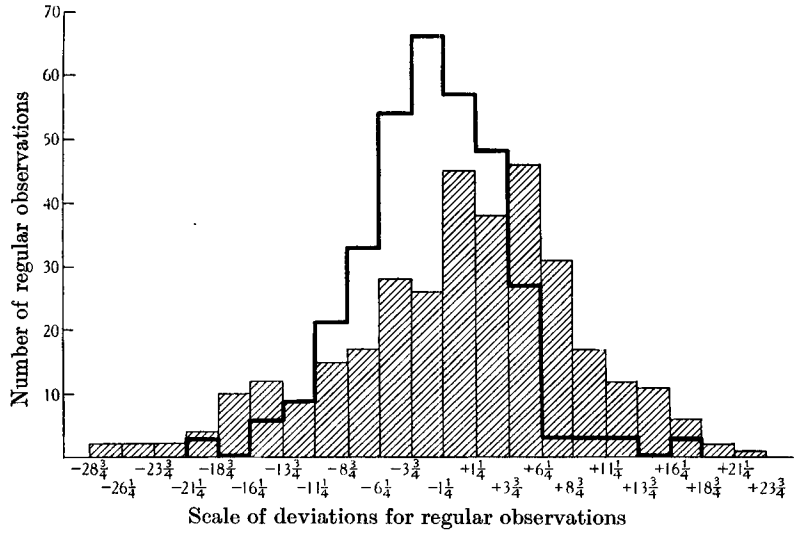

Fig. 3. Distribution of regular observations (shaded) and additional observations (unshaded) on 28th June.

Table II shows the mean heights of the two varieties on the dates in question, computed from the six regular measurements and also from the two additional ones. The change over of the bias as the height increases is well shown. These biases are of very considerable magnitude in relation to the variation of the experimental material. Had the biased observations

been the only ones available, the growth rate between May 31 and June 28 would have been underestimated by as much as 8½ per cent.

The biases are of course of the type that might be expected. When the shoots are low and there is nothing much to be seen except the top leaves there will be a tendency to pick the higher shoots, but when they have come into ear the observer can see the shoots of all heights, and is more likely to select shoots somewhere near the average, omitting both very high and very low values. The strong negative biases of the last set of measurements shows that this selection was not particularly effective in improving the accuracy of the sample.

Table II. *Mean heights and biases in cm.*

Date	Observations	Squarehead's Master			Yeoman		
		Regular observations	Additional observations	Bias	Regular observations	Additional observations	Bias
May 31	Shoot height	51·99	55·16	+3·17	42·99	46·48	+3·49
June 14	Shoot height	83·64	84·77	+1·13	69·48	71·81	+2·33
June 28	Ear height	125·44	122·16	−3·28	112·25	110·09	−2·16

The amount of variation attributable to the various causes can be determined by an analysis of variance. This is given in Table III for the observations of May 31. The method of calculating such tables and a full discussion of their meaning and uses are given by Fisher(1).

The sources of variation are divisible into the six categories shown in Table III. The column headed mean square gives the estimate of the total variance (on a single plot basis) of each

Table III. *Height measurements of May* 31. *Analysis of variance. Single plot basis*

	Degrees of freedom	Sum of squares	Mean square
Between plots:			
Blocks	7	79·1517	11·3074
Varieties	1	293·6225	293·6225
Plot error	7	29·0850	4·1550
Total	15	401·8592	26·7906
Within plots between ½ plots	16	64·8246	4·0515
Within ½ plots between samples	32	86·4963	2·7030
Within samples	316*	367·6096	1·1633

* Two samples contained only four height measurements each.

category. Thus the mean square corresponding to plot error is given as 4·1550; this is the total mean square error per plot from all sources affecting the varietal differences of mean shoot height (i.e. the total plot variation after allowing for differences between blocks and between varieties). Since the mean varietal difference, 8·57 cm.,* is the difference of two

* This differs slightly from the difference of the means of Table II because all samples have been included.

means of eight plot means, the standard error of this difference is $\sqrt{(\frac{2}{8} \times 4{\cdot}1550)}$ or $\pm 1{\cdot}02$.

The plot mean square error can be regarded as made up of a part due to variation of the individual shoots within samples, which is given by the analysis of variance as 1·1633, a part due to variation between samples within half-plots, which is 2·7030 – 1·1633 or 1·5397, a part due to variation between half-plots within plots, which is 4·0515 – 2·7030 or 1·3485, and finally a part due to variation affecting the whole of the height measurements in the plot equally, which is 4·1550 – 4·0515 or 0·1035. It may be mentioned here that only the first two parts can be reduced by more extensive sampling. The last two are due to inherent inequalities in the half and whole plots which would remain if the whole of the material were sampled. Doubling the number of measurements within the sample would approximately halve the first part, and doubling the number of samples with the same number of measurements per sample would approximately halve the second part. If more accurate sampling is required, the best balance between these two alternatives can be decided by taking into consideration the relative work involved in each.

The above estimates of the mean square errors are affected by errors of estimation, and the estimated reductions of error by more extensive sampling are based on the assumption that the possible number of samples is infinite. This is not quite correct, but sufficiently so for our present purpose.

The part of the standard error of the mean height of a single variety due to variation within the samples only is $\sqrt{(1{\cdot}1633/8)}$ or 0·3813, and the standard deviation of a single height measurement about the true mean value of the height of the sample is $\sqrt{(1{\cdot}1633 \times 24)}$ or 5·284. Thus the mean bias, 3·33 cm., in the additional measurements is 0·63 of the standard deviation of a single plant in the distribution from which the selection was made, or 8·74 times the standard deviation due to variation within the samples of the mean of all the height measurements of one variety.

III. Biases in the sampling yields of wheat

Our second example is from the observations for harvest yield. One of the stations, both in 1933 and 1934, took the trouble to harvest and thresh the plots in their entirety for purposes of comparison with the sampling yields. Table IV shows the results.

Table IV. *Yields in cwt. per acre*

Year	Variety	Grain			Straw		
		Whole plots	Sampling	Discrepancy %	Whole plots	Sampling	Discrepancy %
1933	Squarehead's Master	16·8	22·4	+33	39·8	49·9	+25
	Yeoman	24·2	31·9	+32	41·3	57·4	+39
1934	Squarehead's Master	15·6	20·5	+31	32·9	37·5	+14
	Yeoman	12·7	17·2	+35	21·8	28·8	+32
	Victor	20·8	23·5	+13	38·3	42·0	+10

The discrepancies are alarming and, except for Victor in 1934, are remarkably consistent. The similarity between the figures for grain and straw indicates that the ordinary sources of loss between the sampling and full-scale harvesting are unlikely to account for the whole of the differences. The chief of these sources of loss are:

(1) Bird damage after sampling (both before and after reaping).

(2) Shedding of grain.

(3) Carting and threshing losses.

(4) Possibly lower yields of edge rows.

(5) Lower moisture content.

(6) Differences in length of straw.

The total of all these sources of loss undoubtedly at times assumes large proportions. The observers on the spot, however, are of the opinion that in this case the total loss can scarcely have amounted to more than 10 per cent.

The remainder of the discrepancy appears to be due to lack of randomisation in the sampling. The observer at this station, having paced out the requisite number of paces for locating the sample, then took the sample "at random" anywhere in the metre; nor were the four quarter-metres always quite contiguous. That this did introduce considerable bias is confirmed by two points:

(1) The observer admits that he probably tended to select good samples at harvest.

(2) The variety Victor gave a more even crop, so that any bias in the selection of the quarter-metres was of less consequence.

IV. Biases in a crop estimation scheme

This example is provided by a set of sampling tests made in the United Provinces, India, in connection with crop estimation of wheat.

The method of sampling adopted was that of casting a circular hoop, 10 sq. ft. in area, 22 times "at random" within a roughly demarcated area of 1/20th acre. This operation was performed on several experimental farms on fields whose yields were afterwards determined by full harvesting. Twenty-one 1/20th acres were sampled on each field, one of them being the "standard 1/20th acre", which was accurately demarcated and after taking the samples was harvested in full. We can thus compare:

(1) The yields of the standard 1/20th acre given by sampling with that given by full harvesting.

(2) The sampling yield of the standard 1/20th acre with the other 20 sampling yields from the same field.

(3) The yield of the field estimated by both these sampling processes with that given by the full harvesting.

Table V gives the results of the first comparison. The sampling yields are expressed as percentages of the full yields (including the sample). A very considerable bias is shown, the mean percentage being 113·9 with a standard deviation of only 2·97.

Table V. *Standard 1/20th acres on experimental farms. Sampling yields as percentages of yields by harvesting*

122	131	117	115	96
139	118	99	99	109
100	111	93	131	106
96	114	110	96	112
104	135	112	111	99
144	158	119	107	—

Standard deviation of mean ±2·97. Mean 113·90.

Various reasons might be advanced for this bias. The hoop will undoubtedly tend to strike the upstanding parts of the crop and it may perhaps tend to enclose some extra ears; there is also a possibility of loss in the harvesting of the standard 1/20th acre; though this is not likely to be great since it was carried out at the same time. The bulk of the bias, however, is probably due to a tendency, conscious or unconscious, to cast the hoop on the good parts of the crop.

Table VI. *Standard 1/20th acres on circuits. Sampling yields as percentages of yields by harvesting*

Circuit:	Baheri	Bareilly	Nawabganj		Farid pur
			I	II	
	125	96	106	106	117
	125	103	97	90	118
	84	127	98	87	107
	89	87	93	95	83
	88	95	100	98	100
	86	110	110	81	110
	111	124	80	85	109
	62*	98	88	93	121
	85	123	90	131	49*
	100	130	94	89	112
Mean	99·33	109·30	95·60	95·50	108·56
Standard error	±4·54	±4·30	±4·30	±4·30	±4·54

* Gross error? Rejected.

That this bias is by no means constant and is to a certain extent a personal factor is shown by the results obtained from similar measurements on commercial wheat on five circuits. These are shown in Table VI. The differences between the means of the different circuits are just significant, and the means of some of the circuits are certainly significantly different from those obtained by the same process on the experimental farms (Table V).

The second comparison is shown in Table VII. This gives the number of subsidiary 1/20th acres in each field which have a higher sampling yield than the standard 1/20th acre. If the standard 1/20th acre and all the other 1/20th acres were truly random, we should expect all numbers from 0 to 20 to occur with equal frequency. Actually there is a clear bias towards the lower numbers (as shown in Table VII (*b*)), with indications of a reversal of this bias for certain observers. In other words most observers, but not all, have tended to select an

especially good 1/20th acre as standard. Whether this was done deliberately or not is not clear; there do not appear to have been any very specific instructions that the standard 1/20th acre should be "random", or average.

Table VII. *Experimental farms*

(*a*) *Number of subsidiary sheaves* (*out of* 20) *above standard sheaf*

1 5 1½ 3 15½ 13 1 19 14 15 13½ 2½ ½ 3

0 6 5½ 1½ 0 4 2½ 7 1 13 6½ 16 16½ 8 4

A bracket indicates that all the sets of samples were probably taken by one observer.

(*b*) *Classification*

	Observed	Expected
0–5	15½	8·3
6–10	4½	6·9
11–15	5½	6·9
16–20	3½	6·9
Total	29	29·0

In view of the bias in casting the hoop, the sampling estimate of yield is best made by dividing the average yield of the 21 chosen areas of 1/20th acre each by the ratio of the sampling yield of the standard 1/20th acre to its yield by harvesting. When so estimated the yields show no detectable bias if compared with the full harvest of the whole field, available at the experimental stations. The mean of the estimates expressed as percentages of the full yield is 101·72, with a standard error of 2·82. Since, however, about one acre was sampled from each field, and this was in most cases a considerable fraction of the total area of the field, there is no guarantee that such a bias would not appear if much larger areas were so sampled, with a considerably lower number of samples per unit area.

Conclusions

The three examples described in this paper all agree in showing that when any element of personal selection is exercised by the sampler the results are biased; in all three cases a well-designed method of random sampling would have given considerably more accurate results for the same amount of work. The combined evidence of the examples definitely disposes of the claim advanced by some samplers in the case of physical measurements of variable material that they are capable of taking a more representative sample if they are permitted freedom of choice.

In studies connected with mankind the problem of sampling is usually rather different. We do not have to determine the mean height of a group of men by measuring the height of a few picked out from an assembled group. But we do have to make determinations of all kinds on personal characteristics, status, sociological environment, and genealogy by samp-

ling methods, and here again most serious bias may be introduced by improper selection of the material. In a house to house survey, for example, the temptation to avoid houses which are externally non-typical of the district being surveyed is strong; many other motives of a less worthy nature may also operate, such as the difficulty of collecting the required information from certain types of householders, or the mere fact (as in Kiser's example already referred to) that there is no one there to supply the information. Perhaps, however, the most glaring examples of bias in human sampling are provided by postal enquiries or questionnaires. Since only those who are interested in the problem under consideration are likely to trouble to reply, it is clear that such enquiries will be entirely unrepresentative unless very special efforts are made to ensure a complete set of answers.

In view of the evidence, therefore, I do not think we can escape from the conclusion that the only satisfactory method of avoiding bias is for the sampling to be random, whether it be from a wheat field or from a human population. By suitably chosen restrictions and the use of appropriate (possibly complex) sampling units, we can get over the disadvantages on the score of accuracy suffered by purely random samples. In addition, with random sampling we can make the sampling itself provide a really valid estimate of the sampling error with negligible extra labour.

REFERENCES

(1) R. A. Fisher (1932). *Statistical Methods for Research Workers.* 4th edition. Edinburgh: Oliver and Boyd.
(2) C. V. Kiser (1934). "Pitfalls in sampling for population study." *J. Amer. Statist. Ass.* **29**, 250–6.
(3) J. Neyman (1934). "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection." *J.R. Statist. Soc.* **97**, 558–606.