The Statistical and Social Inquiry Society of Ireland.

METHODS OF SAMPLING APPLIED TO IRISH STATISTICS.

By Robert C. Geary, M.Sc.,
Department of Industry and Commerce, Dublin.

[*Read Friday, 27th February, 1925.*]

As the Society is aware, during the recent troubles the system of complete enumeration of agricultural statistics broke down in this country, and since then it has been necessary to rely on estimates derived from a sample of returns received through the post from a very large number of farmers distributed all over the country. It is proposed to change this system this year to a sample of districts, but before doing so it was necessary to investigate the error which the method involved, and this paper is the outcome of that investigation.

I.

A great proposition due to Laplace is fundamental in the theory of simple sampling. Laplace showed that if a variate x is given as the sum of a great number of elemental variates u,

$$x = u_1 + u_2 + \ldots + u_n,$$

then when each u assumes a great number of different values the resulting values of x are distributed according to the Normal Law, under very general conditions governing the values of u, provided that these values are independent.*

Professor Edgeworth in 1904 gave the general conditions which should be satisfied by the elemental frequency distributions in order that Laplace's result should hold.† More recently

*That is to say, if a large number k of samples have been taken, the number of values of x lying between x and $x + d$, when d is small, is

$$\frac{k}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{1}{2}(x-m)^2 / \sigma^2} \cdot d,$$

where m and σ are the arithmetic mean and the standard deviation of all the values of x.

†Camb. Phil. Trans. Vol. XX., Part I., 1904.

in a number of interesting papers Dr. Isserlis showed that when samples of n are drawn from a universe containing N elements then the frequency distributions of the means of these n elements is approximately normal, provided that N was large.* These means are distributed about the universal mean (that is to say, the mean of the N elements) with standard deviation σ_1 given by

$$\sigma_1^2 = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right), \quad \ldots \text{(i)}$$

where σ is the standard deviation for the original distribution.

In the small scale experiments, which will now be described, designed in order to exemplify this result of Dr. Isserlis under local conditions, it should be emphasised that the purpose is not so much to show the striking accuracy of the estimates obtained in the actual experiments as to show rather that the errors were under control. Averages from samples of 100 drawn from 1,000 numbers may possibly differ widely from the average of the 1,000—how widely is going to depend on the original frequency distributions, but the experiments will show that even if the exact error in the average could not have been foreseen the probability of a given error was known *a priori.*

An examination of the publications relating to the agricultural statistics of Ireland showed that from 1847 to 1851 acreage under crops and numbers of live stock were posted by district electoral divisions. There were about 3,600 district electoral divisions in the country. So as to reduce the experiment to reasonable dimensions only the acreages of two crops, wheat and oats, on the first 995 district electoral divisions in the 1851 Report were considered. (The statistics for these divisions were shown in alphabetical order for each poor law union. The poor law unions were also arranged in alphabetical order. Those selected for experiment extended from Abbeyleix to Coleraine, fifty in all.)

So as to obtain random samples of about 100 each, fifteen numbers comprising the last three digits of the area under grass, the area under crops, the total area, the valuation and the population were associated with each district simply as random numbers. The first set of ten samples were obtained by taking all those districts whose "area under grass" showed zero, one, two, up to nine, in the hundreds place. Similarly, the tens and units place of "area under grass," and also the units, tens and

*Journal of the Royal Statistical Society, 1918, pp. 75, *seq.*

hundreds places of the remaining sets of numbers each yielded ten samples. Thus, 150 samples of about a hundred districts each were obtained. Averages of acreage under wheat and acreage under oats per district electoral division were worked for each sample. It should be added that a Powers' Sorting and Tabulating Installation made this experiment but the work of a few hours.

It is necessary to consider the principles underlying the method of selection of these samples. The 150 samples are not independent according to the usual conditions of the general theory. Only fifteen, one derived from each column of numbers, are independent. The fifteen found by taking the zeros on each column are completely independent; so are the averages found by taking the ones, but they are not independent of the averages found by taking the zeros. For instance, a low average found by taking a zero is likely to be followed by a high average on the ones. In spite of this the two series of 150 samples each are distributed in accordance with the Normal Law. This result can be regarded in a number of ways.

In the first place, if the 15 "zero" averages are distributed according to the Normal Law (as far as so few as 15 can be said to have a distribution) around the universal mean with a standard deviation which depends on the frequency distribution of the universe, the 15 one averages likewise, and so on then the aggregate of the 150 averages are going to be distributed according to the Normal Law around the same mean with the same standard deviation. Let us now consider the contrary proposition: the 150 averages selected in the above manner, and therefore not at random, are found to be distributed according to the Normal Law, with the given standard deviation, what can be deduced as to the distribution of randomly selected averages?

Each set of 15 averages have this in common that they are each randomly selected from the universe of all averages (of 100 elements each). The difference between the frequencies of a given average for two series of 150 samples each, even though these two series are not independent, can only be of the order of the probable error of this frequency. Hence the law of distribution of the 150 samples must reproduce the law of the 15. The 150 are found to be normally distributed; therefore each 15 must be normally distributed.

There is yet another remark to be made. Suppose that the first group of random samples shows a distribution whose frequency in a given grade is too low. There will be a very marked tendency on the part of the frequency from the second

group of samples independent amongst themselves but dependent on the first group to make good the deficiency of the frequency in this grade. In other words, there is a very marked tendency in samples selected in this manner towards steadying the aggregate frequency in a given grade. For this reason the distribution found from the 150 samples selected in this manner will be more in accord with the Normal Law than 150 samples selected at random.

This is just what is shown by the two tests I have made. The first, relating to wheat averages, shows that according to Professor Pearson's classic analysis, $P=.831$, which means that if the law of the universe were the Normal Law, then if the experiment of finding the frequency distribution of 150 averages be repeated 1,000 times, no less than 831 out of the thousand distributions would be farther away from the true universal distribution than the 150 averages actually found. Similarly the oats averages show that the value of $P=.984$, which is also significantly high.*

TABLE A.—Showing the frequency distributions of the acreages under wheat and oats on 995 District Electoral Divisions in 1851.

WHEAT.		OATS.	
Acreage (in hundreds).	Frequency	Acreage (in hundreds).	Frequency.
Under 1	610	Under 2	123
1 and under 2	110	2 and under 4	176
2 „ 3	103	4 „ 6	218
3 „ 4	61	6 „ 8	166
4 „ 5	46	8 „ 10	113
5 „ 6	25	10 „ 12	72
6 „ 7	13	12 „ 14	48
7 „ 8	10	14 „ 16	26
8 „ 9	2	16 „ 18	21
9 „ 10	8	18 „ 20	12
10 „ 11	1	20 „ 22	9
11 „ 12	1	22 „ 24	4
12 „ 13	3	24 „ 26	5
13 „ 14	1	26 „ 28	—
14 „ 15	—	28 „ 30	—
15 „ 16	1	30 „ 32	2
Total	995	Total	995

The elemental frequency distributions from which the samples are drawn are going to influence the frequency distributions of the means. Table A. shows the frequency distribu-

*Excluding the single frequency in the highest grade.

tions for the two crops. It will be seen that in each case the distributions were assymetrical. The following were the means and standard deviations and also the expected standard deviations for the averages calculated from the expression—

$$\sigma_1 = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} = \frac{\sigma}{10}\sqrt{\frac{9}{10}},$$

where σ is the standard deviation for the original distribution :—

Crop.	Mean, m	Standard Deviation, σ	Standard Deviation of Average, σ_1
	Acres.	Acres.	Acres.
Wheat	144	194.7	18.46
Oats	676	474.1	44.95

The following table shows the actual distributions of the averages for each crop and also the expected frequencies* on the hypothesis that the averages above obeyed the normal law of error :—

TABLE B.—The actual distributions and theoretical normal distributions of 150 averages of acreages under wheat and oats grown on 100 District Electoral Divisions selected at random† from 995 District Electoral Divisions.

WHEAT.

Deviation from Mean, 144 acres.	Actual Frequencies.	Theoretical Frequencies.
—60 and under —50	1	0.4
—50 „ —40	2	1.8
—40 „ —30	4	5.5
—30 „ —20	13	13.1
—20 „ —10	27	23.2
—10 „ 0	33	30.9
0 „ 10	24	30.9
10 „ 20	23	23.2
20 „ 30	12	13.1
30 „ 40	8	5.5
40 „ 50	2	1.8
50 „ 60	1	0.4
Total	150	149.8

*In all calculations dealing with theoretical frequency distributions, "Tables for Statisticians and Biometricians" were used.

†See text, p. 63.

OATS.

Deviation from Mean, 676 acres.	Actual Frequencies.	Theoretical Frequencies.
—140 and under —120	1	0.4
—120 „ —100	1	1.4
—100 „ —80	4	3.7
—80 „ —60	7	8.0
—60 „ —40	10	14.4
—40 „ —20	21	21.2
—20 „ —0	27	25.8
0 „ 20	28	25.8
20 „ 40	25	21.2
40 „ 60	14	14.4
60 „ 80	6	8.0
80 „ 100	4	3.7
100 „ 120	1	1.4
120 „ 140	—	0.4
140 „ 160	1	0.1
Total	150	149.9

It will be seen that, considering the small number of experiments, the fit with the normal curve is remarkably close and can leave no room for doubt that the law governing the distribution of these averages drawn from such assymetrical original distributions is very close to the Normal Law.

II.

The method of estimate which will now be described may interest the Society, since it is with certain modifications the method actually in use for estimating the acreage under crops and the numbers of live stock in the Irish Free State. We assume that the aggregate measure of an attribute is known in a base year in which a random sample is drawn. The change in the total of the sample at a later period is applied to the known basic measure in order to estimate the measure of the attribute at the later period.

Without any mathematical analysis it is possible to indicate some of the factors upon which the accuracy of this method will depend. In general, each sample drawn from the original material will yield a different ratio. If the original random sample is sufficiently numerous it will at each period yield a good measure of the average, and hence the estimated ratio will be close to the truth. The accuracy of the ratio is therefore

going to depend on the number of the sample; analogy with the method already described would suggest that the error is proportional to $n^{-\frac{1}{2}}$. If now all the elements have altered in the same ratio between the two periods, it is clear that a sample of even one element will yield this proportion. Arguing by continuity, it will be seen that the error in the ratio is going to depend on the amount which the proportions for each element differ from one another. The correlation coefficient r between the measures of the elements on the two dates is a measure of this proportionality, when the measures have altered proportionately in the same direction $r=1$. It may reasonably be assumed therefore that the error will depend also on the factor $(1-r)$.

A mathematical investigation of this method is given in a note at the end of this paper. It will be sufficient to state its results here.

(1) The different ratios resulting from the different random samples selected at the basic period are distributed according to the Normal Law when the sampled number is large.

(2) The mean of all the possible sample ratios is very nearly equal to the unknown ratio for all the elements.

(3) The standard deviations of the distribution of ratios is given by

$$S^2 = 2(1-r)vv'.\left(\frac{m^1}{m}\right)^2\left(\frac{1}{n}-\frac{1}{N}\right), \quad \ldots \text{(ii)}$$

where v and v' are the coefficients of variation (the ratio of the standard deviations to the means) in the two years, m and m^1, the universal means in the two years, n the number sampled and N the number in the universe. It has been assumed that v does not differ much from v'. This appears to be a feature at least of agricultural statistics.

I made some tests of the accuracy of these theoretical results. From the years 1852 to 1872 the volumes relating to the agricultural statistics of Ireland showed the acreages under crops and numbers of live stock in each of the 334 baronies in the country. These were the smallest geographical units for which figures were given, extending over a period of years. The numbers of pigs in the country then as now were subject to violent fluctuations from year to year, and in consequence this statistic seemed suitable for an experiment in sampling; it seemed likely to submit the theory to a severe test.

There are two distinct problems to consider—

(1) How will the ratios, determined from different samples for two given years, differ from the true ratio?

(2) If a sample be selected in the base year, how will its change measure the universal changes over a period of years?

I selected fifteen different series of random numbers—the last three digits of the areas under grass, areas under tillage, total areas, valuations and populations were selected, and each of these fifteen numbers yielded ten dependent samples of about 33 baronies each. The smallness of the sample is rather unfair to the theory which contemplates moderately sized samples. This defect is remedied by a large correlation coefficient, and it will appear that the results are in accordance with the theory. For each sample the ratio of the number of pigs in 1853 to the number in 1852 was calculated. The actual increase for the country as a whole was 6.74 per cent. The functions entering into S (see formula (ii)) were as follows :—

Coefficient of variation,	1852	... .7439
„	1853	... .7402
Coefficient of correlation,	1852-1853	... .9858
Number in sample ...	...	... 33.4
Number in universe ...	...	... 334

Whence S from formula (ii) equals 0.0219.

The following table shows the actual and theoretical frequency distributions. The numbers in the first column show the absolute deviation from the true ratio 1.0674, and the theoretical frequencies in the third column are calculated on the assumption that the distribution was normal and that the standard deviation of the distribution was S.

[TABLE.

TABLE C.—Number of pigs in Ireland, 1852 and 1853.—Showing the actual distribution and the theoretical normal distribution of the ratios between the years 1852 and 1853, derived from 150 samples of about one-tenth each, selected at random* from the 334 Baronies of Ireland.

Deviation from True Ratio, 1.0674	Actual Frequency.	Theoretical Frequency.
—0.08 and under —0.07	—	0.1
—0.07 „ —0.06	1	0.4
—0.06 „ —0.05	1	1.2
—0.05 „ —0.04	4	3.3
—0.04 „ —0.03	7	7.8
—0.03 „ —0.02	17	14.3
—0.02 „ —0.01	25	21.5
—0.01 „ 0	28	26.4
0 „ .01	20	26.4
0.01 „ .02	23	21.5
0.02 „ .03	15	14.3
0.03 „ .04	2	7.8
0.04 „ .05	5	3.3
0 05 „ .06	—	1.2
0.06 „ .07	1	0.4
0.07 „ .08	1	0.1
Total	150	150.0

It will be seen that the correspondence between the actual and the theoretical is very close; that actual experiments showed 32 per cent. of the ratios would lie within 0.01 of the true ratio 1.0674 comparing with the theoretical 35 per cent. Both theory and experiment show that 64 per cent. of the ratios would lie within 0.02 of the true result; in other words, that the odds were two to one against getting a result greater than 1.0874 or less than 1.0474.

The remarks on page 63 regarding the randomness of the 150 samples selected apply to this case also. The freqency distribution shown for these 150 sample ratios is the aggregate of 10 series of 15 experiments each. The samples within each series are independent of one another, but are not independent of the samples in the other series. But assuming as before that there is a law governing each distribution of 15 samples, the law will still be obeyed by the 10 series, and the frequency

*See text, p. 63.

distribution of the 150 samples obtained in this way will for the reasons explained give a closer fit than the distribution of 150 samples selected at random. In Professor Pearson's notation, $p=.809$,* that is to say, if the universal law for the distribution of ratios were the Normal Law and 1,000 absolutely random experiments of 150 averages each be selected and their distributions found, there would be 809 less in accord with the expected frequencies than the actual results found.

At the same time it should be emphasised that even though these results show that remarkable accuracy is attainable by means of a sample of one-tenth, still this is not the best sample of one-tenth upon which to estimate. Much greater accuracy would result from taking the sample of one-tenth completely at random from the smallest practical geographical units, the individual holdings. It will be seen that the coefficient of correlation which plays a fundamental role in the theory between successive years will not be nearly so great for holdings as between baronies, but the number in the sample will be greater, and the latter is generally the stronger force in controlling the error.

From a practical point of view, the variation in the measure of a given sample over a period of years is more interesting than the variations in different samples between two given years, and with this view I selected the baronies whose units place of area showed 0, 1 and 2 in 1852. These numbered 32, 31, and 40 respectively, and for each group I calculated the number of pigs in each year. The following table shows for each sample the number of pigs each year expressed as a percentage of the number for the same sample the previous year. The actual percentages for the country as a whole are also shown.

[TABLE.

* On regrouping the frequencies shown in Table C. two at a time, and neglecting the frequency (2) in the highest grade.

TABLE D.—Number of pigs in Ireland, 1852 to 1874.—The number each year expressed as a percentage of the number the preceding year, actual and estimated from samples of about one-tenth of the Baronies of Ireland.

Year.	Actual	Estimate.		
	334 Baronies	32 Baronies	31 Baronies	40 Baronies
1853	106.74	109.30	106.26	108 56
1854	117.26	117.01	115.88	120.67
1855	87.71	85.49	88.90	87.16
1856	78.00	80.12	78.16	80.71
1857	136.70	134.08	138.35	133.43
1858	112.32	111.15	114.80	113.72
1859	89.78	91.10	89.23	91.52
1860	100.42	100.68	101.62	100.71
1861	86.70	87.89	86.71	86.33
1862	104.74	104.59	101.64	104 88
1863	92.47	91.63	97.47	87.69
1864	99.16	98.04	97.4	96.79
1865	123.38	122.73	125.91	124.37
1866	114.65	114.37	114.48	117.48
1867	82.50	84.13	84.09	81.18
1868	70.40	71.68	69.07	72.51
1869	124.45	121.26	122.08	123.02
1870	135.02	134.00	134.01	135.40
1871	110.96	111.42	113.18	112.49
1872	85.64	86.79	84.34	84.04
1873*	74.90	74.33	78.43	74.70
1874*	105.38	107.35	103.26	104.66

There is therefore a remarkable correspondence between the actual and the sample percentages considering the size of the sample. In the 66 samples there were but two ratios which showed an error of 5 per cent.

The actual number of pigs in Ireland, 1852-'72, and the numbers estimated from the three samples, using the year 1852 as a base, are shown diagramatically. It will be seen that the correspondence in each case is close.

Suppose we test the hypothesis that the 66 percentage errors extending over 22 years are similar in magnitude and distribution to those which would have arisen had 66 samples of the same size been drawn in the year 1852 and the errors in

*Excluding Wexford and Waterford Counties, where in 1873 and 1874 a re-grouping of Baronies took place.

their ratios between 1852 and 1853 measured. The following table is drawn up on this hypothesis—

TABLE E.

Percentage Error in Ratio.	Actual Frequency.	Hypothetical Frequency.
—6 and under —4	1	1.6
—4 „ —2	6	9.2
—2 „ 0	26	22.1
0 „ 2	22	22.1
2 „ 4	9	9.2
4 „ 6	2	1.6
Total	66	65.8

At the risk of bearing too heavily on a single example, it suggests that there is no specially privileged sample which, selected in the base year, will yield better results than all others when used as an index over a period of years.

III.

The first application is naturally to the *fait accompli,* to agricultural statistics. We have seen that the coefficient of variation of the material, the coefficient of correlation between successive years, and the number in the sample play fundamental roles in this theory. The following table shows these functions. I calculated the following from samples selected at random from the returns received last year from farmers in Co. Cork for four of the principal agricultural products :—

TABLE F.—Showing the coefficients of variation and correlation in the years 1923 and 1924 for certain agricultural products, calculated from returns received from Co. Cork.

Product.	Coefficient of Correlation, r	Coefficient of Variation, '23. v	Coefficient of Variation, '24, v'
Pigs*	$.760 \pm .009$	$1.339 \pm .021$	$1.303 \pm .020$
Cattle†	$.909 \pm .005$	$0.834 \pm .018$	$.865 \pm .019$
Sheep*	$.894 \pm .004$	$2.637 \pm .041$	$2.681 \pm .042$
Oats†	$.885 \pm .007$	$1.147 \pm .025$	$1.186 \pm .026$

*Calculated from 961 returns.
†Calculated from 480 returns.

Apart from their application to the theory of sampling, these functions have a certain interest in themselves. In the first place, the correlation coefficient for pigs is significantly less than for the other three products. The number of pigs on individual holdings varies considerably from year to year. All such movements which make the ratio for individual farms differ considerably will tend to depress r. The economic reason for this irregular movement on individual holdings is probably due to the fact that big changes can be brought about with safety and economy much more rapidly in pigs than in sheep, and in sheep than in cattle, simply because the sow is much more prolific than the ewe and the latter than the milch cow, and also because of the relative lengths of the gestation periods and the periods from birth to slaughter or export.

It has already been remarked that the coefficient of variation for agricultural products does not change much from year to year. It will also be seen that the coefficient of variation for sheep is much greater than for the other products shown. One of the reasons for this is that on a large proportion of holdings no sheep are raised at all. These mathematical functions therefore are significant of certain aspects of the economics of the products to which they refer.

We will now examine the accuracy of the estimate based upon the two methods. Table G. shows the standard deviation expressed on a percentage of the actual result for samples of three different sizes; these correspond roughly to the numbers of returns used in estimating the acreage under crops and number of live stock in the Irish Free State last year: 2,500 an average for the county, 16,000 an average for the province, and 64,000 returns used for the whole country. The number of returns used was roughly one-seventh of the holdings in the Irish Free State. The numbers in the compartments are calculated on the assumption that each sample is one-seventh of the "universe" from which it is drawn; the numbers shown mean that the probability of a percentage error as great as or greater than that shown is 1 in 3, of an error twice as great 1 in 22, or three times as great 1 in 370.

[TABLE.

TABLE G.—Showing the standard deviations expressed as percentages of the mean value, resulting from two methods of estimate, distinguishing these standard deviations for samples of different sizes.

Product.	Size of Sample.					
	2,500		16,000		64,000	
	Meth. I.	Meth. II.	Meth. I.	Meth. II.	Meth. I.	Meth. II.
Pigs	2.49	1.81	0.98	0.72	0.49	0.36
Cattle	1.66	0.73	0.65	0.29	0.33	0.14
Sheep	5.12	2.42	2.02	0.96	1.01	0.48
Oats	2.29	1.14	0.91	0.45	0.45	0.23

Method I.—Simple Sample.
Method II.—Method of Ratios.

The only estimates published are those for the whole country, based on 64,000 returns, calculated by Method II., namely, the method of ratios. It will be seen that the odds are 21 to 1 against an error of 0.7 per cent. for pigs, 0.3 per cent. for cattle, 1.0 per cent. for sheep, and 0.5 per cent. for oats.

The question of error by the method of ratios appears in practically all questions relating to index numbers of production and prices where these depend fundamentally on arithmetic averages. This application to the most familiar and homely of all indices, the cost of living index number, may now be considered. The general principal of calculation adopted in all countries is the same: the index represents the price of fixed quantities of certain goods which cost 100 at some previous (base) date. The current prices used are the weighted average prices derived from prices returns received from a large number of retailers.

Ultimately the calculation of the aggregate index depends on the index numbers for individual commodities. The index number for a commodity is the quotient of the sum of its current prices divided by its prices at the base period. But the number of prices received from retailers, though large, is still small compared with the number of retailers in the country. One may ask what would the index be had other retailers rendered returns.

For every commodity high and low prices are current, and, furthermore, through all price movements the high price district is likely to remain high and the low price district to remain

low. There will therefore be a positive correlation between prices ruling at different dates. The coefficient of variation for the prices of the commodity will be relatively low. The theory shows that these two factors will ensure that the error in the index or ratio is small. A further examination of the question would be interesting; it would probably reveal that, from the point of view of cost of living, the index number would not be made appreciably more accurate even if twice or three times the large numbers of returns used in its compilation were used. Without divulging any official secrets it may interest the Society to know that a collateral set of returns of retailers' prices received from Post Office officials and used in the Department of Industry and Commerce for the purpose of checking its own prices showed a cost of living index at the last enquiry of 195.2, comparing with the official 195.4.

IV.

It has been necessary to omit from consideration the more delicate applications of the theory: the accuracy attainable, for instance, in the classification of a total according to a limited number of attributes. How far, for instance, may the proportions at different ages established from a random sample of ten thousand persons be regarded as the true proportions for the country as a whole? How far may the proportions in a number of industrial groups derived from a sample of ten thousand workers be regarded as the correct proportions for all workers in the country? Striking results may be obtained in the application to these and similar questions provided the samples satisfy the two *criteria* of all sampling—

(1) The samples are random.

(2) The number sampled is large.

The first condition is readily satisfied in laboratory or small scale experiments. The application of sampling to life-sized economic problems is much more difficult. It seems paradoxically to be true that the golden rule of random sampling is to leave nothing to chance.

In order to make the sampling statistic a success it is, if not essential, at least of the first importance that the distribution of the questionnaire and the method of selection of the sample should be rigidly controlled by a central authority. A card index showing, if not all the individuals in the country, at least the householders, would be the great ideal and is not so far from realisation. It might easily be made in a census year. As late as 1918 an up-to-date list of all farmers in the country was

in existence, and the Factory and Workshops Register, also quite up to date, contains the name of every factory and workshop in the most general sense of the word. The Trade Boards Lists supplement the Factory and Workshop List in many particulars. All these lists could be card indexed, revised from time to time and numbered according to some random system.

The selection of the sample will then become completely automatic.

The possibility of the application of the method to the fundamental statistics of the country raises a number of considerations. In spite of its accuracy and cheapness, no statistician is an exponent of sampling *sans phrases*. We have seen that in dealing with certain agricultural statistics great accuracy may be attained to, that an error of one-half of one per cent. in an estimate based on a sample of one-tenth in a simple total for the country is very remotely possible. Suppose that this statistic is made up of 32 county totals also estimated on samples of one-tenth, and consider two counties, Cork containing 40,000 agricultural holdings and Carlow containing only 6,000 holdings, all Ireland containing 570,000 holdings. Then an error of one-half of one per cent. is equiprobable with an error of 3.8 per cent. for Cork and 9.7 per cent. for Carlow. If the country total is made up of the individual totals for 155 poor law unions, then the error of one-half per cent. is equiprobable with an error of 12.6 in the figure for the poor law union. These applications result from the applications of the Law of Error in its simplest form: that the accuracy of a sampled total is inversely proportional to the square root of the number in the sample. As the ultimate unit decreases the relative accuracy of its total also decreases. Before the sample method is applied to a specific problem two questions must therefore arise—

(1) In what detail are the statistics to be presented?

(2) What degree of accuracy is required in the detail?

Dealing with the first question, it is clear why the statistician is unwilling to take the responsibility of a drastic change in the manner of presentation of the most fundamental of existing statistics or in initiating new statistics why he is so particular to err on the right side in the question of detail. He must be essentially conservative. His existing statistics for a given year are but a part and incomplete if they are not comparable with those for previous years. He is also a custodian for the future: his statistics of to-day may only assume a significance twenty years hence. One may say that his universe,

like Einstein's, has a time dimension. A lessening of detail in a certain direction (possibly in order to extend in others) means the closing of a chapter. To expect him to cut his detail and adopt a more economic classification seems to be begging the question.

With regard to the second question, had not the purpose of the meticulous accuracy in statistics been a little misunderstood? It is a truism that the last three digits in a five-figure total have little significance, and yet perhaps the last three figures have entailed as much work in their compilation as the first two. A sample method may conceal a change of one per cent. in either direction in a year. This change, taken by itself, rarely conveys anything to the economist and still less to the mind of the man in the street. Now if this total be examined in detail this one per cent. may be very significant. It may be the resultant of a number of opposite tendencies: a decrease of five per cent. in the total of a single county steadied by slight changes in the opposite direction for the other counties. Or a steady total for a single county may conceal a definite tendency in its poor law union areas.

There does not appear to be any general answer to these questions. A tentative general solution would appear to be that a sampling *régime* will never be substituted for the great periodic enquiries. Even from the point of view of the sample a complete enumeration is a necessary base to the best methods of estimate.

In certain cases it may be possible to effect considerable economies even in the most important statistics without a sacrifice of detail. An example may be taken from trade statistics. A characteristic of the imports into this country is the comparatively large proportion entered in small consignments and whose aggregate value is comparatively insignificant. This is a persistent feature of several of the most important commodities. If all the consignments of these commodities whose value is £10 and over be included and an estimate, say, of one-third, made for the remainder by considering only those consignments whose value was under £10 entered on, say, the Monday and Thursday of every week, a very close estimate of the imports of this commodity will be obtained. The value will be close, partly because the total value involved in the estimate is small and partly because the number of consignments is large.

But perhaps, considering everything, the sample may find its most important applications in the bye-ways of economics where absolute accuracy is not required, in those questions which must at present remain obscured in the realm of surmise because they are not deemed sufficiently important to be

examined by a cent. per cent. enumeration. Whether the enumeration is to cover one-tenth, one hundredth, or one thousandth even, of the material will depend on the accuracy required in the results.

It only remains for me to express my thanks to my chief, Mr. John Hooper, whose suggestions and criticisms at all stages of this paper have been invaluable to me.

THE DISTRIBUTION OF RATIOS.

The measures of N elements in two successive years are

$$m + u_1,\ m + u_2,\ \ldots\ m + u_N,$$

and

$$m' + u_1',\ m' + u_2',\ \ldots\ m' + n'_N,$$

where m and m' are the means in the respective years. The real ratio is therefore m'/m. An estimate of this ratio is found by taking a random sample of n elements. This ratio, a is given by

$$a = \sum_n (m' + u_i') \,/\, \sum_n (m + u_i) = \frac{m'}{m} \cdot (1 + x') \,/\, (1 + x)$$

where

$$x = \sum_n u_i \,/\, nm \quad \text{and} \quad x' = \sum_n u_i' \,/\, nm'.$$

Let $b = (1 + x')/(1 + x)$. Now assume that the number N in the universe and the ratio n/N are such that the distributions of the simple averages derived from n elements are approximately normal in both years.* The distribution of x and x' are normal. It is also known that the joint error (x, x') is distributed on the normal surface of error with a coefficient of correlation ρ.† The distribution of the ratio a, when n is fairly large, will be nearly normal.

In fact x is a quantity of order $n^{-\frac{1}{2}}$ when the coefficient of variation is a small finite number. Knowing this coefficient, and since the distribution of x is normal, a number n can be found so that any defined proportion of samples drawn from the original N elements will give an x less than 1 in absolute value. The fraction b can then be expanded in positive powers of x,

$$b = 1 + x' - x + x^2 - xx' - x^3 + x^2x' \ \ldots \qquad \text{(i)}$$

retaining only powers up to the third of (x, x'). Indicating arithmetic averages of large numbers of samples by square brackets,

*Isserlis, *Journal of the Royal Statistical Society*, 1918, pp. 75, *seq.*

†Edgeworth, *Camb. Phil. Trans.*, vol. xx., Part 1, 1904.

the first moments of the distribution of b are given by

$$\left.\begin{aligned}\mu_1 &= [b] = 1 + \sigma^2 - \sigma\sigma'\rho \\ \mu_2 &= [(b - \mu_1)^2] = \sigma^2 - 2\sigma\sigma'\rho + \sigma'^2 \\ &\quad + \sigma^2\{8\sigma^2 - 16\rho\sigma\sigma' + (3 + 5\rho^2)\sigma'^2\}\end{aligned}\right\} \quad \ldots \text{(ii)}$$

retaining only the first significant powers and using the relations

$$[x^4] = 3\sigma^4 \; [x'^4] = 3\sigma'^4$$

$$[x^3x'] = 3\rho\sigma^3\sigma', \quad [x^2x'^2] = (1 + 2\rho^2)\sigma^2\sigma'^2, \quad [xx'^3] = 3\rho\sigma\sigma'^3,$$

where σ and σ' are the standard derivations of x and x'.

A characteristic of agricultural statistics is that the coefficient of variation changes but slightly from year to year. Transform (ii) on the assumption that $\sigma' = \sigma\,(1 + \epsilon)$, where ϵ is small. Then (ii) become

$$\left.\begin{aligned}\mu_1 &= 1 + \sigma^2\{(1 - \rho) - \epsilon\rho\} \\ \mu_2 &= 2\sigma^2(1 - \rho)(1 + \epsilon).\end{aligned}\right\} \quad \ldots \qquad \text{(iii)}$$

It will now be proved that the distribution of b is approximately normal. In fact, over the greater portion of its range b may be accurately represented by $b = 1 + x' - x$. To find what is the effect of this assumption, suppose that $n = 1{,}000$ and the coefficient of variation of the original material is 2. The probability that x will be as great as or greater than $3 \times 2 / \sqrt{1000}$ is 0·0027 or 1 in 370. The first term neglected, $x(x - x')$, is, therefore, practically certain to be less than 1/5 of the last term retained.

But $(1 + x' - x)$ is itself the mean of n quantities drawn from a universe whose measures are

$$1 + \frac{u_1'}{m'} - \frac{u_1}{m}, \quad 1 + \frac{u_2'}{m'} - \frac{u_2}{m}, \quad \ldots \quad 1 + \frac{u'_N}{m'} - \frac{u_N}{m},$$

and, therefore, is normally distributed under the usual conditions governing n and N.*

It will now be proved that ρ, the coefficient of correlation between all the averages of n elements in the two years, equals the coefficient of correlation r between the measures of the original N elements in the two years.

---

* The values of $\beta_1 = \mu_3^2 / \mu_2^3$ and $\beta_2 = \mu_4 / \mu_2^2$, where μ_3 and μ_4 are the third and fourth moments of the distribution of b, are

$$\beta_1 = 16\sigma^2\{(1 - \rho) - (1 + \rho)\epsilon\},$$
$$\beta_2 = 3 + 3\sigma^2\{(14 - 10\rho) - 10(1 + \rho)\epsilon\},$$

retaining only terms in σ^2 which is of order n^{-1}. The next terms are of order n^{-2}. When, for example, $n = 1000$, the coefficient of variation of the original material 2, $\rho = \cdot 9$ and $\epsilon = \cdot 01$, then $\beta_1 = \cdot 004$ and $\beta_2 = 3 \cdot 059$. If the distribution were normal, these functions should be zero and 3 respectively. For the example given in the text relating to the numbers of pigs raised on Irish baronies, $\beta_1 = \cdot 001$ and $\beta_2 = 3 \cdot 188$, which are still sufficiently near the normal values.

In fact, ρ is given by

$$n^2 {}^N C_n \rho \sigma_1 \sigma_1' = \Sigma (u_1 + u_2 + \ldots + u_n)(u_1' + u_2' + \ldots + u'_n),$$

where σ_1 and σ'_1 are the standard deviations of the averages in the two years and summation on the right-hand side is extended to ${}^N C_n$ terms of which the first is given resulting from the selection of the u's with the corresponding u''s taken n at a time. Then

$$n^2 {}^N C_n \rho \sigma_1 \sigma_1' = {}^{N-1}C_{n-1} \Sigma u_i u'_i + {}^{N-2}C_{n-2} \sum_{i \neq j} \Sigma (u_i u_j' + u_j u_i')$$

$$= N \cdot {}^{N-1}C_{n-1} ss'r - N \cdot {}^{N-2}C_{n-2} \cdot ss'r, \ldots \quad \text{(iv)}$$

where s and s' are the standard deviations of the u's and u''s. Now it has been shown* that

$$\sigma_1 = \frac{s}{n} \sqrt{n - \frac{n(n-1)}{N-1}} \quad \text{and} \quad \sigma_1' = \frac{s'}{n} \sqrt{n - \frac{n(n-1)}{N-1}}.$$

In equation (iv), dividing across by $n^2 \cdot {}^N C_n$, and using these results, we find $\rho = r$.

The result may now be definitively stated:—

Given N elements whose measures in successive years are

$$m + u_1, \; m + u_2, \; \ldots \; m + u_N,$$

and

$$m' + u_1', \; m' + u_2', \quad m' + u'_N,$$

measured from their means m and m'. Then the ratios

$$\sum_n (m' + u'_i) \; / \; \sum_n (m + u_i)$$

are distributed about a mean which equals m'/m *approximately according to the normal law with standard deviation* S *given by*

$$S^2 = 2(1 - r) vv' \cdot \left(\frac{m'}{m}\right)^2 \cdot \left(\frac{1}{n} - \frac{1}{N}\right),$$

where v and v' are the coefficients of variation of the original material and N and n are large.

Corollary.—The question arises: For a given value of n will the method of simple sampling or the method of ratios, where it may be applied, yield the better results?

By the method of simple sampling the averages are distributed normally round a mean m' with standard deviation σ_1'; by the other method the ratios are distributed normally round a mean m'/m with standard deviation S given by

$$S^2 = 2(1 - r) \sigma_1'^2 \cdot \frac{1}{m^2},$$

where it is assumed that the coefficients of variation are equal in the two years.

In general, the errors by the second method will be less than by the first so long as

$\sigma_1' > mS$ or upon reduction so long as $r > \frac{1}{2}$.

* Isserlis, *loc. cit.*

DIAGRAM showing, for each of the Years 1852-1872, the number of Pigs in Ireland, actual and estimated, the latter derived from samples selected from the 334 Baronies of Ireland.