SAMPLING TECHNIQUE FOR OBTAINING NUMBER OF COVERED WORKERS UNDER STATE UNEMPLOYMENT COMPENSATION LAWS

Harry J. Winslow *

Proposals to modify State unemployment compensation laws usually involve questions on benefit costs. For the State legislatures, which must secure such information at a minimum of time and cost, a small sample from the wage-record file will suffice to answer many questions. A small representative sample properly planned yields better results than a larger sample poorly planned and requires very little more time to obtain than one improperly selected. A survey of all the records would ordinarily be both unnecessary and uneconomical. In a recent sample, drawn for the purpose of measuring the number of workers with wage credits within a calendar year, special techniques for random sampling were developed and should be helpful in planning samples for measuring wage characteristics of covered workers.

The administrative procedures developed by State employment security agencies for keeping records of workers' earnings make it difficult, except by carefully planned sampling procedures, to obtain an unduplicated count of the number of workers who have had some earnings in covered employment during a year. In most States the record of a worker's annual earnings can be obtained from one or more punch cards or wage slips filed for each quarter for which the worker has had earnings. If a worker receives wages from more than one employer during a quarter, the file will contain a wage-record card for each employer. The total number of workers who have had some earnings in covered employment during a year can be obtained by counting all the workers with cards or slips in the wage-record file, counting of course only one slip per worker. This method is, however, laborious and time-consuming, particularly in a large agency in which a large volume of wage-record cards are received during the year.

A study was made of the Maryland wage-record file for 1938 to develop a sampling procedure which would be simple to carry out, require a minimum amount of clerical and machine time, and provide sufficient data to make a reliable estimate of the number of workers with wage credits and to detect the error in the estimate. The extent to which this goal was achieved is shown by the amount of time required for pulling, processing, and refiling the cards, and by the accuracy of the results. The sample was drawn in November 1939. A record was kept of the time required for all the hand and machine operations, and adequate statistical data were tabulated to give not only the end results but also to estimate the accuracy of the sample. Additional data were obtained in order that measures of the number of workers with wage credits might be made by two independent methods, to detect, if possible, any bias that might have occurred.

Summary of Results

Three random samples were selected independently of each other. Each sample was approximately 0.8 percent of the total universe. Two methods were devised for obtaining the estimated number of workers with wage credits. These methods were so designed that the number of workers estimated by each would move in opposite directions from the true number of workers if the sample were biased. This procedure provided a check on the effectiveness of the device used for eliminating the bias which results from obtaining too large a proportion of workers with a large number of wage-record cards per worker.

For the first method, the following procedure was adopted. The total number of cards in the file was obtained by measuring the number of inches of cards in the file, then sampling to determine the average number of cards per inch, and multiplying the total number of inches by the average number of cards per inch. The number of workers with wage credits was obtained by

*Bureau of Employment Security, Research and Statistics Division. The author wishes to acknowledge the cooperation and assistance of the Maryland Unemployment Compensation Board in making available statistical data for use in the preparation of this article. The article presents the findings of the first of a series of sampling experiments designed to determine efficient and satisfactory specifications for random sampling of wage-record files. It does not represent official recommendations of the Division.

determining from the sample the average number of cards per worker and then dividing the total number of cards by this figure. The average number of workers obtained from the three samples by this method was 427,000, with an error of ±1.7 percent. This error was calculated for 95 percent fiducial limits.[1] The relative error within each individual sample was ±2.5 percent.

In the second method, the number of workers with wage credits was obtained by dividing the total earnings reported in the annual report for 1938 by the average annual earnings estimated from the three samples. The number of workers calculated by this method was 425,000, with an error of ±2.7 percent. The error within each individual sample was ±4.6 percent. Errors were calculated for 95 percent fiducial limits.

Precautions were taken to eliminate any bias that might result from selecting too high a proportion of workers with large numbers of wage-record cards. The agreement between the number of workers computed by each method was sufficiently close to indicate that there is little likelihood of a large bias in the sample toward workers with a large number of cards. Although this study does not offer positive proof that there is no bias, it does give favorable evidence that the selection of the samples was quite random. Reasons for this belief are discussed below in detail.

The Maryland Wage-Record Files

Since the Maryland wage-record files segregate the wage records for the current year from the preceding year, the records needed for the study were immediately available and the sampling was thereby greatly facilitated. Maryland has a uniform benefit year from April 1 to March 31. The base period is the calendar year preceding April 1 of the current year. As a result, in April a large number of initial determinations of benefit rights are made daily. After the first few weeks, the number of determinations drops off rather rapidly and becomes very small in the latter part of the year. At the time this sample was selected—November 1939—few initial determinations were being made. Since the cards in the wage-record file are drawn at the time the determination is made and are then replaced, few cards are out of the file at any one time, especially in the latter part of the year. This refiling has a slight effect on the order of the cards in the file. At the beginning of the year the cards are filed by machine and are in perfect order, but as the year progresses the possibility increases that some of the cards will be out of place in the file. Even so, little evidence was found to indicate that many cards had been misfiled.

The 1938 wage-record files contained about 32 sections of 10 drawers each. Each drawer contained 2 trays. A full drawer held approximately 5,000 cards. However, many of the drawers were not full. In those drawers that were considered full, the number of inches of cards in each tray varied from 22 to 29. A few of the trays were only partially filled. In all, excluding cards for railroad workers, there were 630 trays, containing approximately 1,601,000 cards.

The file consisted of two main sections, a numeric file (by social security account number) and an alphabetic file. The numeric file contained cards for (a) workers to whom social security account numbers had been assigned in the State of Maryland, (b) those who had received their account numbers outside the State, and (c) workers with wage credits under State coverage who had initially received account numbers from the block issued under the railroad retirement system. The last two groups were relatively small.

The alphabetic file represented cards for workers for whom no social security account number was reported. As soon as an account number was obtained for these workers, their cards were shifted to the numeric file. In the alphabetic file, the cards were filed alphabetically by the worker's surname only and then filed under surname by employer account number. There was no way to determine the number of separate records for individuals having the same initials or given name. Since, fortunately, the alphabetic file was small—12 trays of cards only—the error due to duplication was well below the magnitude of the sampling error for the principal result. The method of filing used in the alphabetic file made it difficult to select cards for individual workers. Drawings from this file could have been

[1] The percentage error represents the range on either side of the true mean of the universe that will include 95 percent of the cases. In other words, the odds are 19 to 1 against obtaining a mean that will lie outside these limits by chance alone. Inasmuch as the mean of the universe is unknown, and no factors that would cause a systematic error or bias have been detected in the measurement of the cards, the odds are against the 427,000 workers differing from the actual number of workers by more than ± 1.7 percent.

omitted entirely without impairing the accuracy of the data.

At the time the sample was drawn, wage-record cards were being pulled at the rate of about 1,000 a day. The cards remained out of the file no longer than 2 days. Since the average number of cards per worker was 3.75, wage records for no more than 600 workers were out of the file at any one time. Scattered throughout the file were cross-reference cards. Whenever wage records for the same individual were found under different account numbers, the cards were placed under one number and a cross-reference was inserted for the other number. The total number of these cross-reference cards was small. Records of workers covered by railroad unemployment insurance were in a separate part of the numeric file and were therefore readily excluded from this study on workers covered by the State law. The general set-up provided an almost ideal sampling arrangement for determining the average number of cards per worker.

Method of Drawing Sample and Time Required

Wage records for five workers were drawn at random from every full tray of cards. If the tray was half full, cards for three workers were drawn; if less than half full, for only one worker. The cards which were pulled were spaced fairly equally along the tray. The trays were broken at five different places without actually measuring the space between breaks. To eliminate the bias that would have resulted from random pulling of the cards of the first worker, the cards of the second worker were always drawn. For example, when the clerk broke the file, instead of pulling the first card at this break, he would draw the cards for the next following number. In case the second worker's card happened to be a cross-reference, this card was pulled from the file as though it were a regular account and placed in a separate pile.

If the cards for the first worker had been drawn, a definite bias would have occurred in the sample in favor of the workers with the largest number of cards. Since the number of cards per worker varied greatly, the space occupied by individual workers' wage records was unequal. In breaking the file at random, the probability of breaking the file for a worker whose wage records occupied a wide space in the file was much greater than breaking a file at a worker whose wage records occupied a narrow space. By taking the second lot of cards following the break, this bias was, for the most part, eliminated. The probability of a second lot containing a large or a small number of cards was practically the same.

Two of the samples were drawn independently by two different individuals. The third sample was pulled by three different individuals because of changes in staff on duty. The time required for pulling the three samples from the numeric file was: first sample, 7 hours, 3 minutes; second sample, 5 hours, 49 minutes; third sample, 8 hours, 23 minutes. The time required for sampling the alphabetic file was: first sample, 44 minutes; second, 41 minutes; third, 45 minutes.

The total time required for drawing all three samples from both files was 23 hours, 25 minutes. The time for refiling the cards was 67 hours, 18 minutes. This time was longer than it should have been, because no guide cards were placed in the trays where the cards had been pulled. If file guides had been used, the time for refiling would have been considerably reduced. A conservative estimate of refiling under these conditions would be 45 hours.

In order to tabulate the data, summary cards were punched for each worker, with the following information: social security account number, number of cards per worker, and total annual earnings, From this information the following tables were prepared:

(1) Distribution of workers by number of cards per worker and by type of account number or other identification.

(2) Distribution of annual wages by amount and by type of account number or other identification.

(3) Number of workers earning $3,000 or more and total amount of individual earnings in excess of $3,000, in groups of 200 workers each, arranged in account-number sequence.

(4) Number of cross-reference cards by type of account number.

(5) Identification of the last worker in each group of 200 workers in each sample, by account number.

The time required for machine work was: sorting, 18 hours, 15 minutes; tabulating, 19 hours, 10 minutes; miscellaneous,[2] 7 hours, 45 minutes. The total time spent on machine operations for

[2] Includes verification of runs, collating and balancing, wiring machines, and incidental clerical work.

all three samples was 45 hours, 10 minutes. The total time required for both clerical and machine work was 135 hours, 53 minutes, or an average per sample of 45 hours, 18 minutes. Had file guides been used, it is estimated that the average time per sample would have been 37 hours and 20 minutes.

The total number of workers drawn was 3,357 for the first sample, 3,316 for the second, and 3,291 for the third, making a total of 9,964. The average amount of clerical and machine time required for each worker included in the sample was 49 seconds.

In order to estimate the time required for pulling a sample from a similar arrangement of files, an approximate figure can be obtained by multiplying the total number of workers in the proposed sample by the average time per worker. For the most part this figure will be an overstatement, because all the tabulations included in this study are unnecessary, and the time for refiling the cards can be cut down.

Number of Cards in Files

In some States the number of wage cards filed during the year is known. The total number of cards punched for the year 1938 in the Maryland agency was 1,725,000. This figure includes cards for railroad workers.

Certain difficulties are inherent in estimating the number of cards by the procedure used in the first method described. The number of cards per inch will vary according to the proportion of new or used cards in the drawer, humidity conditions, pressure on the cards, and the position of the drawer in the file cabinet. To ensure reasonably uniform pressure on all cards, the same clerk made all the measurements. It was found that the number of cards per inch in a tray in an upper drawer varied from that in a lower drawer because of the difference in leverage which could be applied at the time of measurement. In spite of this difficulty, it is believed that this method of estimating the number of cards in a file gave fairly accurate results. The total number of inches of cards was 11,757, of which 570 were accounted for by cards for railroad workers. Since several independent sets of measurements were not taken, it is impossible to estimate the over-all error in these figures.

In order to determine the average number of cards per inch, six batches of cards measuring 6 inches each were selected at random throughout the file. The cards in each batch were run through a sorter and counted. The number of cards per batch is shown below:

Batch number	*Number of cards*
1	862
2	869
3	869
4	868
5	864
6	863

The average number of cards per inch was 144.3 ± 0.4 percent. The error is expressed for 95 percent fiducial limits.

Table 1.—*Number of workers represented in 3 samples drawn from 1938 wage-record files of Maryland Unemployment Compensation Board, by number of cards per worker*

Number of cards per worker	All samples					Sample 1					Sample 2					Sample 3				
	Total	Maryland	Out-of-State	Railroad	Alphabetic	Total	Maryland	Out-of-State	Railroad	Alphabetic	Total	Maryland	Out-of-State	Railroad	Alphabetic	Total	Maryland	Out-of-State	Railroad	Alphabetic
Total number of cards	37,402	.....	....	....	....	12,636	.....	....	....	....	12,326	.....	....	....	....	12,440	.....	....	....	....
Total number of workers	9,964	8,984	757	62	161	3,357	3,026	255	21	55	3,316	2,984	256	21	55	3,291	2,974	246	20	51
1	1,298	924	268	20	86	455	324	93	9	29	450	328	92	4	26	393	272	83	7	31
2	976	777	148	9	42	306	236	56	1	13	331	270	42	4	15	339	271	50	4	14
3	869	759	97	3	10	282	246	29	2	5	288	246	38	1	3	299	267	30	0	2
4	5,453	5,240	188	15	10	1,861	1,705	57	5	4	1,795	1,720	67	5	3	1,797	1,725	64	5	3
5	637	602	27	5	3	215	208	6	0	1	213	196	13	3	1	209	198	8	2	1
6	294	277	11	5	1	94	85	6	3	0	92	90	1	0	1	108	102	4	2	0
7	121	114	5	1	1	36	33	3	0	0	37	35	0	1	1	48	46	2	0	0
8	107	101	4	1	1	34	33	1	0	0	41	37	2	1	1	32	31	1	0	0
9	50	46	2	2	0	18	15	2	1	0	17	16	0	1	0	15	15	0	0	0
10–14	73	65	2	1	5	21	19	0	0	2	30	26	0	1	3	22	20	2	0	0
15–19	36	30	4	0	2	14	12	1	0	1	9	7	1	0	1	13	11	2	0	0
20–24	22	22	0	0	0	8	8	0	0	0	5	5	0	0	0	9	9	0	0	0
25 and over	28	27	1	0	0	13	12	1	0	0	8	8	0	0	0	7	7	0	0	0

When cross-reference cards were pulled, they were placed apart from the cards with wage records. The number of cross-reference cards per sample was as follows:

Sample number	Total number of cards	Cross-reference cards	
		Number	Percent of total
All samples	37,402	295	0.79
1	12,636	99	.78
2	12,326	116	.94
3	12,440	80	.64

The number of cross-reference cards thus constituted a very small proportion of the total. In fact, this number is well within the range of sampling error resulting from measuring the average number of cards per inch and the average number of cards per worker. For all practical purposes, this count could be excluded entirely from the estimate without affecting the accuracy of the end results, since the total number of cards per inch is reduced only to 99.2 percent when the correction for cross-reference cards is made. When the average number of cards per inch, corrected for cross-reference cards, is multiplied by the total number of inches of cards in the file, the total number of wage cards is 1,601,000±0.4 percent. Since cards for railroad workers are excluded from this estimate and the percentage error is calculated for 95 percent fiducial limits, the error in the measurement of total number of cards in the file is very small. However, the statistical estimate of the error for 95 percent fiducial limits does not take into account the error that might have occurred in this measurement.

As previously stated, about 2,000 cards were out of the file at the time these measurements were taken. This number is well within the range of accidental error and can be disregarded in estimating the total number of cards in the file. Regardless of the difficulties in measuring the average number of cards per inch, when proper care is taken it can be done quickly and accurately. The percentage of error in this measurement is much less than the percentage of error that occurs in measuring the average number of cards per worker.

The time required for measuring the cards in the sample was not recorded, because in most States the total number of cards that have been filed is known. Even if this time were included in the total clerical and machine time, the average time per covered worker would not be increased appreciably. In all likelihood this average figure is sufficiently overestimated to include the time necessary for measuring the files and the average number of cards per inch.

Estimate of Number of Workers

A distribution of workers by number of cards per worker was tabulated for each sample (table 1). The number of cards per worker varied from 1 to 35. For each sample the modal number of cards per worker was 4. The mean number of cards per worker and the error in the mean for 95 percent fiducial limits is as follows:

Sample number	Average number of cards per worker	Standard deviation	Percent of error in the average
All samples	3.75	2.31	±1.2
1	3.76	2.40	±2.1
2	3.72	2.23	±2.1
3	3.78	2.30	±2.1

The standard deviation in each sample is almost as large as the mean. However, the error in the mean is relatively small. For each sample the error was ±2.1 percent, whereas the error for all three samples combined was ±1.2 percent. Since the number of workers in each sample is about 0.8 percent of the total number of workers in the universe, the proportion for all three samples combined is about 2.4 percent of the universe. It is readily seen, therefore, that improvement in accuracy does not increase in direct proportion to the increase in the size of the sample. To reduce the error of the average number of cards per worker 45 percent, the size of the sample must be increased 200 percent.

In order to estimate the number of workers by a different method, a distribution of workers by annual earnings was tabulated (table 2). The average annual earnings for each sample and the error in this average for 95 percent fiducial limits are shown below:

Sample number	Average annual earnings	Standard deviation	Percent of error in the average
All samples	$896	$1,213	±2.7
1	872	1,064	±4.7
2	888	1,318	±4.6
3	928	1,246	±4.5

The standard deviation for each sample is greater than the mean itself. The error for each sample is ±4.7, ±4.6, and ±4.5 percent, respectively; for all three combined, ±2.7 percent. It should be noted that the error in the mean for 95 percent fiducial limits is more than twice as great as the corresponding error for the average number of cards per worker. The calculation of the number of workers with wage credits by this method is, therefore, much less accurate than the estimate based on the average number of cards per worker. This difference is to be expected, because individual annual earnings of workers spread over a greater range than the number of cards per worker.

The highest individual annual earnings obtained in any one sample was $100,000; the next highest was $65,300. Both these earnings were much higher than any others drawn. Federal income-tax reports for the State of Maryland show that the total number of workers earning over $50,000 a year is a very small fraction of 1 percent of the total number of workers in the State. The chance of drawing a worker with annual earnings of over $50,000 in a sample as small as these is very remote, and it is best to exclude such records in determining the average earnings of all workers. If they are left in the sample and the average annual wage is computed, a less accurate estimate of the mean is obtained. In the calculation of the average annual earnings, the worker with $65,300 annual earnings was not included in sample 1, and the worker with $100,000 annual earnings was not included in sample 3. Some idea of the exaggeration that would have occurred if the highest individual earnings in sample 1 and sample 3 had been included may be obtained by observing the earnings of the next highest worker in each of these samples. The earnings of the next highest worker in sample 1 was $22,200 against the highest annual earnings of $65,300. The annual earnings of the next highest worker in sample 3 was $21,000, as compared with the highest earnings of $100,000. The proportion of workers earning more than $10,000 in these samples was 0.2 and 0.3 percent, respectively.

An average of annual wages does not represent the average annual rate for total man-years of employment, but is the average annual-earning rate per worker, regardless of the amount of unemployment an individual may have experienced during the year. It corresponds roughly to the average annual earnings obtained from pay rolls and employment under the old-age and survivors insurance program and not under the State unemployment compensation laws. The latter average more nearly corresponds to an average full-time annual-earning rate per man-year of employment.[3]

Estimates of the number of workers with wage credits in 1938, as previously stated, can be obtained (1) by dividing the total number of cards in the wage-record file by the average number of cards per worker; and (2) by dividing the total wages reported in the annual report for 1938 by average annual earnings as estimated from wage-record cards.

[3] Total man-years of employment for this rate are obtained by averaging monthly volumes of employment for a whole year. Monthly volume of employment is defined in employment security statistics as the number of workers employed within the pay-roll period ended nearest the last day of the month.

Table 2.—*Number of workers represented in 3 samples drawn from 1938 wage-record files of Maryland Unemployment Compensation Board, by annual-earning group*

Annual-earning group	All samples					Sample 1					Sample 2					Sample 3				
	Total	Maryland	Out-of-State	Railroad	Alphabetic	Total	Maryland	Out-of-State	Railroad	Alphabetic	Total	Maryland	Out-of-State	Railroad	Alphabetic	Total	Maryland	Out-of-State	Railroad	Alphabetic
Total amount of earnings	$8,927,000	......	...	...	....	$2,928,000	......	...	...	....	$2,946,000	......	...	...	....	$3,054,000	......	...	...	....
Total number of workers	9,964	8,984	757	62	161	3,357	3,026	255	21	55	3,316	2,984	256	21	55	3,291	2,974	246	20	51
Less than $100	1,662	1,266	254	20	113	575	435	91	11	38	553	424	85	9	35	534	407	78	9	40
100-499	2,363	2,099	211	16	37	776	692	67	4	13	829	735	72	6	16	758	672	72	6	8
500-999	2,561	2,416	123	12	10	879	829	41	5	4	827	777	42	5	3	855	810	40	2	3
1,000-1,499	1,750	1,668	78	4	0	590	559	30	1	0	567	544	22	1	0	593	565	26	2	0
1,500-1,999	865	832	31	1	1	287	278	9	0	0	284	271	12	0	1	294	283	10	1	0
2,000-2,499	347	316	31	0	0	122	113	9	0	0	122	112	10	0	0	103	91	12	0	0
2,500-2,999	165	153	12	0	0	54	50	4	0	0	55	48	7	0	0	56	55	1	0	0
3,000-3,999	120	108	12	0	0	35	33	2	0	0	43	39	4	0	0	42	36	6	0	0
4,000-4,999	33	31	2	0	0	14	12	2	0	0	7	7	0	0	0	12	12	0	0	0
5,000-9,999	72	71	1	0	0	18	18	0	0	0	21	20	1	0	0	33	33	0	0	0
10,000 and over	26	24	2	0	0	[illegible]	7	0	0	0	8	7	1	0	0	11	10	1	0	0

Estimates of the number of workers with wage credits obtained by these two methods are shown below:

Sample number	Sample as percent of total	Workers estimated by average number of wage cards		Workers estimated by average annual earnings	
		Number	Percent of error	Number	Percent of error
All samples	2.4	427,000	±1.7	425,000	±2.7
1	.8	426,000	±2.5	436,000	±4.7
2	.8	431,000	±2.5	428,000	±4.6
3	.8	424,000	±2.5	410,000	±4.5

Errors for these two estimates are for 95 percent fiducial limits. The error for the number of workers calculated from the average number of wage cards was obtained by adding the error in the average number of cards per worker to the error in the estimate of the total number of cards in the wage-record file. The error in the second estimate is due entirely to the error in average annual earnings per worker. There is no way of determining the error in the annual report on the amount of wages in covered employment for the State of Maryland. Since this amount was tabulated from a 100-percent sample, the error may be presumed to be much smaller than any of the sampling errors shown here.

Variation in the number of workers for each of the three samples in the second estimate is much greater than in the first. Likewise, the errors in the second estimate are almost twice as great. This difference occurs because annual earnings are not as homogeneous a characteristic as the average number of cards per worker.

Estimates of the number of workers with wage credits resulting from the two methods do not differ significantly. It is reasonable to assume, therefore, that there is little or no bias in the method of sampling used. The only bias that might occur—selecting workers with the greatest number of cards—was practically eliminated by selecting cards of the second worker when drawing the sample. If there had been a bias in the measure of the average number of cards per worker, it is possible that this bias would be in the same direction but would not affect average annual earnings to the same extent. Workers with the largest number of wage-record cards are usually workers who have the least stable employment and the lowest wage rates. If these workers occurred in the sample in a greater proportion than they occurred in the universe, the average annual earnings would be little affected, whereas the average number of cards per worker would be much greater than the true average number of cards per worker in the universe. In this event, the bias would cause the two sets of estimates to diverge. The estimate of the number of workers by average number of cards per worker would be too small, whereas the estimate based on average annual wages would be nearer the true value, and the difference between the two estimates would be significant if this bias were sufficiently large. The estimates of the number of workers with wage credits by the two methods do not differ significantly in any of the samples. This fact is substantial evidence that there is little or no bias toward the worker with the greatest number of cards.