UNITED STATES DEPARTMENT OF AGRICULTURE (RESTRICTED)
BUREAU OF AGRICULTURAL ECONOMICS
WASHINGTON, D. C.

IST-41
9-20-49

AGRICULTURAL ESTIMATES: In-Service Training

- Sampling With Probabilities Proportional to Size -

We are about to start an experimental sampling study for which we will later examine the mathematical formulas in detail. The data used are North Carolina county acreages of Peanuts Grown Alone for all Purposes in 1940 and 1945, as listed in the 1945 U. S. Census. Each member of the class will draw an independent sample of 15 counties; one county will be selected in each Crop Reporting District, with the exception of District 3 where 8 counties will be selected instead of one. The counties will be selected within Districts with probabilities proportional to size; the 1940 peanut acreage being the pertinent measure of size. State estimates of 1945 acreages will be made from the various samples of 15 counties each by computing percent change from 1940 for the sample counties by Districts and adding the District estimates to get the 1945 State total. The necessary information for drawing the sample and expanding it is given in the following table. In addition, every member of the class is provided with a different part of Yates' tables of random numbers.

Table 1. - Peanuts Grown Alone for All Purposes - NORTH CAROLINA

DIST.	COUNTY	1940 Acres	1945 Acres	1940 Cumulative Total
1	Alleghany	0	0	0
	Ashe	0	0	0
	Avery	0	0	0
	Caldwell	40	18	40
	Surry	9	2	49
	Watauga	0	0	49
	Wilkes	5	6	54
	Yadkin	22	2	76
Northern Mountain		76		
4	Buncombe	0	1	0
	Burke	88	24	88
	Cherokee	0	0	88
	Clay	1	0	89
	Graham	0	0	89
	Haywood	0	0	89
	Henderson	0	0	89
	Jackson	0	4	89
	McDowell	1	20	90
	Macon	0	1	90
	Madison	1	0	91
	Mitchell	0	0	91
	Polk	16	20	107
	Rutherford	181	75	288
	Swain	0	0	288
	Transylvania	0	0	288
	Yancey	0	0	288
Western Mountain		288		

DIST.	COUNTY	1940 Acres	1945 Acres	1940 Cumulative Total
2	Alamance	46	23	46
	Caswell	7	1	53
	Durham	2	6	55
	Forsyth	5	41	60
	Franklin	25	4	85
	Granville	20	1	105
	Guilford	37	15	142
	Orange	6	0	148
	Person	3	0	151
	Rockingham	19	1	170
	Stokes	0	0	170
	Vance	5	0	175
	Warren	688	671	863
Northern Piedmont		863		
5	Alexander	43	58	43
	Catawba	408	177	451
	Chatham	8	7	459
	Davidson	44	41	503
	Davie	19	18	522
	Iredell	86	21	608
	Lee	7	7	615
	Randolph	21	17	636
	Rowan	221	179	857
	Wake	91	20	948
Central Piedmont		948		
8	Anson	58	51	58
	Cabarrus	129	71	187
	Cleveland	161	256	348
	Gaston	128	160	476
	Lincoln	205	75	681
	Mecklenburg	100	86	781
	Montgomery	34	455	815
	Moore	31	201	846
	Richmond	70	486	916
	Stanly	37	0	953
	Union	44	4	997
Southern Piedmont		997		

DIST.	COUNTY	1940 Acres	1945 Acres	1940 Cumulative Total
3	Bertie	32,232	36,890	32,232
	Camden	326	227	32,558
	Chowan	10,874	11,617	43,432
	Currituck	62	542	43,494
	Dare	1	0	43,495
	Edgecombe	20,248	25,342	63,743
	Gates	11,504	11,350	75,244
	Halifax	37,355	36,410	112,602
	Hertford	22,510	22,534	135,112
	Martin	19,786	24,535	154,898
	Nash	3,764	4,345	158,662
	Northampton	37,125	39,708	195,787
	Pasquotank	443	509	196,230
	Perquimans	7,381	10,161	203,611
	Tyrrell	538	791	204,149
	Washington	6,035	6,506	210,184
Northern Coastal		210,184		
6	Beaufort	917	3,038	917
	Carteret	1,527	1,507	2,444
	Craven	314	530	2,758
	Greene	161	1,134	2,919
	Hyde	7	44	2,926
	Johnston	262	206	3,188
	Jones	1,283	1,683	4,471
	Lenoir	528	776	4,999
	Pamlico	32	5	5,031
	Pitt	5,709	15,685	10,740
	Wayne	473	218	11,213
	Wilson	344	932	11,557
Central Coastal		11,557		
9	Bladen	6,741	9,223	6,741
	Brunswick	4,758	3,748	11,499
	Columbus	2,081	3,994	13,580
	Cumberland	1,002	901	14,582
	Duplin	2,166	2,390	16,748
	Harnett	16	30	16,764
	Hoke	12	161	16,776
	New Hanover	315	736	17,091
	Onslow	6,623	3,219	23,714
	Pender	3,915	3,050	27,629
	Robeson	408	1,132	28,037
	Sampson	827	935	28,864
	Scotland	14	358	28,878
Southern Coastal		28,878		
State Total		253,791	290,428	

The process of drawing a sample of counties is simple. For example, to select the sample county from District 1, take a random number from 1 to 76. Locate the first cumulative total in the column of 1940 cumulative peanut acreages that contains this random number. The county corresponding to that number is the sample county. It will be noted that any random number from 1 to 40 would select Caldwell, a number from 41 to 49 would select Surry, and so on. In District 4, the county would be selected in the same way except that a random number from 1 to 288 is needed for the selection.

District 3 is of particular interest here because 8 counties are to be drawn. The first random number within the required range is certain to select a county, but any following draw may possibly hit a county already selected. When that happens it is necessary to continue drawing random numbers until we reach the quota of 8 different sample counties. For purposes of analysis later it is important that a record be kept of each drawing so that every member of the class can go to the data later and determine which counties were hit by each drawing and the order in which they were hit. For example, Harold Walker obtained the following sample of counties on 19 draws:

County:	Draw:
Bertie	1, 4, 9, 13, 15
Gates	12
Halifax	2, 8, 11
Hertford	7, 16, 17
Martin	6, 10
Nash	19
Northampton	3, 5, 18
Washington	14

The numbers after the names of the counties tell which random draws hit each county as well as the order in which they hit. Bertie County, for example, was selected on the first draw, but it was hit again on draws 4, 9, 13, and 15. The 19th draw hit the last county needed to complete the sample of 8. We will make use of that record later.

It is clear that on the first draw this method of sampling gives each county a probability of selection exactly proportional to the 1940 peanut acreage. When more than one county is drawn from a stratum as in District 3, the probability of a random draw hitting a county is also exactly proportional to the 1940 acreage. But we may hit a county that was already selected on an earlier draw, and we are not taking a county more than once. That limitation has the effect of disturbing the proportionality between probability of selection and 1940 peanut acreage in the county. Later on we will look into this matter more closely; for the time being, we will proceed just as though such a limitation did not exist. We proceed with our computations just as though every county from District

3 in our sample of 8 was selected with a probability exactly proportional to the 1940 peanut acreage. Under that assumption the estimate of the District ratio of 1945 peanut acreage to the 1940 peanut acreage computed from the sample is $\bar{R} = 1/8\ (R_1 + R_2 + R_3 + R_4 + R_5 + R_6 + R_7 + R_8)$ in which the 8 R's are individual ratios computed separately for each of the 8 counties in the sample.

It may seem strange that we use the straight average of the individual county ratios without weighting each one by the 1940 peanut acreage in the county. The answer is that in this kind of sampling the weighting is automatically occurring by the representation of the different size counties. A simple example makes this clear. Suppose we have a stratum containing only 2 kinds of counties, one set having 40 acres of peanuts per county in 1940 and the other 5 acres of peanuts per county. Suppose further that we have 1000 counties of each kind in our population. Now suppose that the counties with 40 acres of peanuts in 1940 show a 10 percent increase in 1945 while the others show no increase. The universe then has the characteristics shown in Table 2.

Table 2. -- Hypothetical Universe Containing only 2 Kinds of Counties

1940 Peanut acreage per county	Number of Counties	Total 1940 Peanut acreage	Ratio of 1945 acreage to 1940 acreage
40	1000	40,000	110
5	1000	5,000	100

The 1945/1940 ratio for the entire stratum is

$$\frac{(40{,}000)(110) + (5{,}000)(100)}{45{,}000} = 108.9$$

It is clear that the first group of counties carries 8 times as much weight as the second in determining the percent change for the entire stratum.

Now assume that we draw a sample of 180 counties out of this stratum with probabilities proportional to the 1940 peanut acreages. As the counties of the first kind are 8 times as large as those of the second, we expect to draw 8 large counties for every small county that is drawn. The composition of our sample should, therefore, be as shown in Table 3.

Table 3. -- Expected Composition of Sample of 180 Counties Selected with Probabilities Proportional to 1940 Peanut Acreages.

1940 Peanut Acreage per County	Counties in Sample	1940 Peanut Acreage in Sample	Ratio of 1945 Acreage to 1940 Acreage
40	160	6400	110
5	20	100	100

UNITED STATES DEPARTMENT OF AGRICULTURE (RESTRICTED) IST-42
BUREAU OF AGRICULTURAL ECONOMICS 9-27-49
WASHINGTON, D. C.

AGRICULTURAL ESTIMATES: In-Service Training

Introduction to Mathematics of Sampling with Probabilities Proportional to Size

In this lesson we will formulate a mathematical model that can be used for studying samples of the kind taken in District 3. It is not the only mathematical model that could be used and it may not be the best; in fact, we will later on consider a slightly different one in connection with a different method of expanding the sample. At the moment we do not know which of those methods of expanding the sample is the better, one of the reasons for working through this experimental sampling problem is to get some information on that point. The mathematical model that we will consider at the moment involves regarding the population of 16 counties in District 3 as appearing in 16 different strata with 1 county per stratum. We will let the symbol P_i represent the probability of selecting a county from the i-th stratum on a single draw. As a single draw is certain to hit one of the 16 counties we must have

$$\mathop{S}_{i=1}^{16} (P_i) = 1.$$

In our particular problem the P_i are proportional to the 1940 peanut acreages; they can be computed simply by dividing the 16 individual county acreages by the total acreage in the District. This gives the following results, using the same order in which the counties are listed in IST - 41:

$P_1 = 0.15335$	$P_5 = 0.00000$	$P_9 = 0.10709$	$P_{13} = 0.00211$
$P_2 = .00155$	$P_6 = .09633$	$P_{10} = .09414$	$P_{14} = .03512$
$P_3 = .05174$	$P_7 = .05473$	$P_{11} = .01791$	$P_{15} = .00256$
$P_4 = .00029$	$P_8 = .17774$	$P_{12} = .17663$	$P_{16} = .02871$

To open the discussion we will first consider the simpler problem of how this model would behave if all of the P_i were equal to each other; in other words, if every county were given an equal chance of coming into the sample as in ordinary random sampling. Representing that probability by P, we would have $P = 1/16$, because on a single draw we would have one chance in 16 of hitting any particular county. Suppose now we see what happens when a single county is selected at random from the list of 16 and each of the 16 counties is given an equal chance of being selected. First of all, we know that one county will be selected. We can write the number 1 opposite the name of that county and record a zero opposite the name of each of the other 15 counties. These numbers are the observed frequencies with which each of the 16 counties appears in that particular sample. We could try the experiment over again, and again write the number 1 opposiet the name of the county selected and zeros opposite all of the others. If we were to continue this sort of experiment indefinitely, we would find that opposite the name of each of the 16 counties we would have the digits 0 and 1 appearing a large number of times, with 1/16 of those digits being 1 and 15/16 of them being o. The distribution of those ones and zeros for any county would represent a random arrangement. Consequently, we can say that the average or expected number of times each of the 16 counties in the universe appears in a sample of one county is 1/16 which, of course, is equal to P.

Now suppose that we repeat the entire experiment but take 2 counties in our sample each time. By putting 1 opposite each of the 2 counties selected and 0 opposite the 14 not selected each time, we would find that when we ended the experiment 2/16 of the digits opposite each county would be 1 and 14/16 would be 0. Consequently, we can say that in a random selection of a sample of 2 counties, the average or expected number of lines each of the 16 counties appears is equal to 2/16 or 2P. In general, we say that the expected number of times each county in the universe appears in a sample of n counties is equal to nP. In this particular example we find that if we take $n = 16$, the expected number of times each county appears in the sample is $16P = 1$. That is merely another way of saying that when the number of counties in the sample is as large as the number of counties in the universe, we are certain to have every county in the universe included in the sample.

As there is only one county in each of the 16 strata, as we picture the situation, we can say that the expected number of times each of the 16 counties appears in a sample of n also represents the expected fraction of that stratum that appears in a sample of n. For example, a total sample of 8 counties would take an expected fraction of $8P = 8/16 = \frac{1}{2}$ of the counties from each of the 16 strata. We will represent this expected fraction for a sample of n counties by $\hat{p}$. Now we can write the equations

$$\hat{p} = nP$$

or

$$P = \hat{p}/n$$

In any sample of n counties we will represent the observed fraction taken from a stratum by p_i This observed fraction will be either 1 or 0, depending upon whether or not the corresponding county was selected. With this concept in mind we can write any formula involving data for a sample of n counties in terms of the entire universe. For example, if X_i represents the 1940 peanut acreage in a county, the per-county average for a random sample of n counties from the population of 16 can be written

$$\bar{x} = \frac{p_1X_1 + p_2X_2 + p_3X_3 + --- + p_{16}X_{16}}{n}$$

If we are talking about a sample of 8 counties, 8 of the p_i will be equal to 1 and the remaining 8 will be equal to zero. Therefore, we would have only 8 terms other than zeros in the expression $\overset{16}{\underset{i=1}{S}} p_iX_i$ and the formula reduces to the ordinary expression for the arithmatic mean of 8 values of X_i. But writing the formula for $\bar{x}$ in the form given above has quite a few advantages for purposes of mathematical analysis. For example, it is a simple matter to prove that $\overline{x}$ is an unbiased estimate of the corresponding population mean. The expected value of any of the p_i is given by

$$E(p_i) = \hat{p} = nP$$

Hence the expected value of $\overline{x}$ is given by

$$E(\overline{x}) = P(X_1 + X_2 + X_3 + \text{---} + X_{16}) =$$

$$1/16(X_1 + X_2 + X_3 + \text{---} + X_{16})$$

which is the exact mean for all counties in the District.

Now consider what happens when the probability of selecting a county is proportional to the 1940 peanut acreage; that is when the P_i have the values given earlier in this paper. If we stick to the mathematical model we have been talking about, the expected fraction taken from the i-th stratum in a sample of n counties is given by, $\hat{p}_i = nP_i$. We are now ready to show that according to this model the proper average 1945/1940 ratio for estimating the percent change in peanut acreage for District 3 is given by the straight average of the ratios for the individual counties in a sample of n. The straight average of the ratios for a sample of n is given by

$$\overline{R} = \frac{p_1R_1 + p_2R_2 + \text{---} + p_{16}\,R_{16}}{n}$$

where the p_i are either zero or unity (n of them are unity). Using the same reasoning that was followed earlier in this lesson, the expected value of this average is

$$E(\overline{R}) = \frac{\hat{p}_1R_1 + \hat{p}_2R_2 + - - - + \hat{p}_{16}\,R_{16}}{n} =$$

$$\frac{nP_1\,R_1 + nP_2R_2 + \text{---} + nP_{16}\,R_{16}}{n} =$$

$$P_1R_1 + P_2R_2 + \text{---} + P_{16}\,R_{16}$$

But since the P_i are proportional to the 1940 acreages and the sum of the 16 values of P_i is equal to 1, the quantity $\underset{i=1}{\overset{16}{S}} P_i\,R_i$ is simply the weighted average 1945/1940 ratio for the 16 counties the District with the 1940 county acreages serving as the weights. This proves the important principle that <u>if our mathematical model were vigorously correct</u>, the <u>straight</u> average of the individual county ratios is actually an unbiased estimate of the weighted average ratio for the District. Several of our people have been worried about this matter of working with the straight average of these ratios in the sample data; the proof just given is the mathematical way of showing how the method of drawing the sample automatically takes care of the weighting. This proof, together with the discussion in IST-41, should clear the matter up. In future lessons we will show that the model with which we are working has some defects; we will examine those defects and try several different methods of overcoming them. We will find that theoretically the straight average of the individual county ratios requires an adjustment in order to make it an unbiased estimate of the weighted average District ratio.

AGRICULTURAL ESTIMATES: In-Service Training

- Sampling With Probabilities Proportional to Size (Continued) -

In IST-42 we proved that, if the probability of selection for the i-th county in a sample of n is equal to P_i (where P_i is proportional to the 1940 peanut acreage), the straight average of the n values of the ratios R_i for the individual counties actually is an estimate of the weighted average ratio for the District. However, we warned that there was a defect in our model. It was evident in drawing your samples that some counties were hit by random numbers more than once, even though those counties were not used more than once, and that this generally happened to the larger counties in the District. Obviously if the larger counties tend to be taken out of the universe on the early draws, all following draws can only make selections from among the smaller counties that are left. As a result we find that only for a sample of 1 county is the probability of selection for the i-th county exactly equal to P_i; for samples of n greater than 1 the probability of selection for the i-th county will not be exactly equal to P_i. In fact, as n approaches 100 percent of the universe, the true probability approaches 1/N as in ordinary random sampling with equal probabilities.

We will now investigate a method of correcting for that discrepancy that Hendricks suggested about 2 years ago but that has not been tested up to this time. Later on we will compare the results with those from other methods that have been proposed for dealing with the problem.

The problem is brought to the fore rather forcefully when we consider the formula given in IST-41 for computing the expected fraction of counties taken from the i-th stratum in a sample of n:

$$\hat{p} = nP$$

When we apply this formula in situations where P varies from county to county, we have

$$\hat{p}_i = nP_i$$

When we do this we often find that nP_i comes out greater than unity, which of course, is nonsense. That actually happens with the first county in District 3; in a sample of 8 counties we would have:

$$\hat{p}_i = (8)(0.15335) = 1.22680$$

The problem reduces to finding the correct probability P_i' which must be substituted for P_i when n is greater than 1. P_i' clearly depends upon the total sample size, for it has the value P_i when $n = 1$ and the value $1/N$ when $n = N$. That exact probability can be computed but the computations would be extremely tedious except in a simple laboratory exercise where we might consider a sample of 2 or 3 drawn from a small population of 5 or 6. The arithmatic would be prohibitive in a practical problem.

Hendricks has proposed an approximation that we will test in this class. The approximation is based on the plausible assumption that the expected number of counties selected from the i-th stratum on a single draw is equal to the product of P_i and the number remaining in that stratum from previous draws. On the first draw we have $\hat{p}_i = 1\ P_i$. On the second drawing we get $(1 - P_i)P_i$ making a total for the 2 draws equal to $\hat{p}_i = 1\ P_i + (1 - P_i)P_i = 1 - (1 - P_i)^2$. It can be shown that for t drawings we have

$$\hat{p}_i = 1 - (- P_i)^t$$

It should be noted that t refers to the number of random numbers that have been drawn and not to the number of counties selected. After drawing t random numbers we expect $\hat{p}_i$ counties to be selected from the i-th stratum. Hence, the total number of counties selected from a universe of N counties by drawing t random numbers is

$$n = \underset{i=1}{\overset{N}{S}} \hat{p}_i = N - \underset{i=1}{\overset{N}{S}} (1 - P_i)^t$$

In our problem this means finding the value of t that will give us

$$16 - \underset{i=1}{\overset{16}{S}} (1-P_i)^t = 8$$

or

$$\underset{i=1}{\overset{16}{S}} (1-P_i)^t = 16-8$$

Such an equation can not be solved for t by any simple, straight-forward algebracic process. But it can be solved without too much difficulty in the present case by letting t take on the successive values 1, 2, 3, 4 ---, until we reach a value that makes the sum of the $(1 - P_i)^t$ equal to 16 - 8. For our data in District 3, the required value of t is slightly over 18 (approximately 18.27). From the relation $\hat{p}_i = 1 - (1 - P_i)t$ we can compute the correct value of $\hat{p}_i$ for each stratum for our sample of 8. It should also be noted that the correct probability of selection for each of the 16 counties in a sample of 8 is given by

$$P_i' = \hat{p}_i/8$$

These exact probabilities and the ratio P_i/P_i' are shown in the following table for each county in District 3.

County	P_i	P_i'	P_i/P_i'
Bertie	0.15335	0.11901	1.2885
Camden	.00155	.00349	.4441
Chowan	.05174	.07762	.6666
Currituck	.00029	.00066	.4394
Dare	.00000	.00000	1.0000
Edgecombe	.09633	.10533	.9146
Gates	.05473	.08028	.6817
Halifax	.17774	.12149	1.4630
Hertford	.10709	.10919	.9808
Martin	.09414	.10444	.9014
Nash	.01791	.03514	.5094
Northampton	.17663	.12140	1.4549
Pasquotank	.00211	.00473	.4461
Perquimans	.03512	.05993	.5860
Tyrrell	.00256	.00572	.4476
Washington	.02871	.05157	.5567

The ratios in the last column of the table are of particular interest because they are used to get an unbiased estimate of the average ratio $\bar{R}$ for District 3. In a sample of 8 counties, the ratio for each county in the sample should be multiplied by the corresponding value of P_i/P_i' for that county. The value of $\bar{R}$ is then obtained by dividing the sum of the 8 products by 8. It is easy to prove that this gives an unbiased estimate of the 1945/1940 District ratio. Writing the formula for $\bar{R}$ in terms of all 16 counties, we have

$$\bar{R} = \frac{p_1(P_1/P_1')R_1 + p_2\,(P_2/P_2')\,R_2 + \text{---} + p_{16}\,(P_{16}/P_{16}')R_{16}}{8}$$

$$E\,(\bar{R}) = \frac{\hat{p}_1\,(P_1/P_1')\,R_1 + \hat{p}_2\,(P_2/P_2')\,R_2 + \text{---} + \hat{p}_{16}(P_{16}/P_{16}')R_{16}}{8} =$$

$$\frac{8\,P_1'\,(P_1/P_1')R_1 + 8P_2'\,(P_2/P_2')R_2 + \text{---} + 8\,P_{16}'\,(P_{16}/P_{16}')R_{16}}{8} =$$

$$P_1 R_1 + P_2 R_2 + --- + P_{16} R_{16}.$$

Each member of the class is requested to recompute the average 1945/1940 ratio for his sample of counties from District 3 by this method and use this new value for the District of $\bar{R}$ to get a better estimate of the 1945 peanut acreage in North Carolina. The results will be compared with the results obtained previously when the straight average of the 8 county ratios from District 3 was used to represent the District.

This method of weighting has some interesting properties. If only one county had been selected from District 3, we would have $t = 1$ and $P_i' = P_i$. If all 16 had been included in the sample we would have $t = \infty$ and $P_i' = 1/16$. Our formulas thus give exact results at the 2 extremes. In between, we are dealing with approximations; but there is good reason for believing that the approximations are close to the truth. The members of the class may be interested in knowing how the 16 values of P_i' were computed. The first step was to compute the 16 values of $(1 - P_i)^t$ for values of $t = 1, 2, 3, - - - -$, each time getting the sum of those 16 quantities. For $t = 18$ that sum was 8.04208. For $t = 19$ the sum was 7.88404. As we want the sum to be exactly $16 - 8 = 8$, we know that the exact value of t is somewhere between 18 and 19. These data are shown below:

County	$(1-P_i)^{18}$	$(1-P_i)^{19}$	Difference
1	0.04997	0.04231	0.00766
2	.97246	.97095	.00151
3	.38433	.36444	.01989
4	.99478	.99449	.00029
5	1.00000	1.00000	.00000
6	.16150	.14594	.01556
7	.36308	.34321	.01987
8	.02953	.02428	.00525
9	.13018	.11624	.01394
10	.16870	.15282	.01588
11	.72232	.70938	.01294
12	.03025	.02491	.00534
13	.96269	.96066	.00203
14	.52544	.50699	.01845

County	$(1-P_i)^{18}$	$(1-P_i)^{19}$	Difference
15	.95491	.95247	.00244
16	.59194	.57495	.01699
Total	8.04208	7.88404	.15804

If we multiply each of the differences in the last column of this table by the factor 0.04208/0.15804 = 0.26626, and subtract the result from the corresponding value of $(I - P_i)^{18}$ we should get 16 values of $(1 - P_i)^t$ that add up to exactly 16 - 8 = 8. We do not need to know the exact value of t corresponding to this adjustment, but it obviously has a value somewhere in the neighborhood of 18.27. These adjusted values of $(1-P_i)^t$ are shown below, together with the estimates of $\hat{p}_i$ and estimates of the "exact" probability, P_i', of selection for each of the 16 counties in a sample of 8.

County	$(1-P_i)^t$	$\hat{p}_i = 1 - (1 - P_i)^t$	$P_i' = \hat{p}_i/8$
1	0.04793	0.95207	0.11901
2	.97206	.02794	.00349
3	.37903	.62097	.07762
4	.99470	.00530	.00066
5	1.00000	.00000	.00000
6	.15736	.84264	.10533
7	.35778	.64222	.08028
8	.02814	.97186	.12149
9	.12647	.87353	.10919
10	.16447	.83553	.10444
11	.71887	.28113	.03514
12	.02883	.97117	.12140
13	.96215	.03785	.00473
14	.52053	.47947	.05993
15	.95426	.04574	.00572
16	.58742	.41258	.05157
Total	8.00000	8.00000	1.00000

The values of P_1' shown above are the values that appear in the second column of the table on page 3.

The numerical work shown above may seem to be excessive. But it should be noted that when we are drawing samples from fairly large populations it is possible to use short-cuts in estimating the appropriate value of t. Most of the sampling units likely to be encountered in our work have frequency distributions with respect to size that can be represented fairly well by a Pearsonian Type III curve:

$$dF = \frac{a^b}{\Gamma(b)} e^{-aX} X^{b-1} dX$$

in which $a = \bar{x}/\sigma_x^2$

$$b = (\bar{x}/\sigma_x)^2$$

If we assume this sort of frequency distribution it is fairly easy to show that

$$\frac{N}{N + v^2 t} = \left(\frac{N - n}{N}\right)^{v^2}$$

in which N = number of units in the population

n = number of units in the sample

$v = \sigma_x / \bar{x}$ = coefficient of variability of size of units in the population.

It is instructive to see how this formula works in District 3, where N = 16 for our sample of n = 8. The squared coefficient of variability of the 16 1940 peanut acreages is 0.9836. The equation becomes

$$\frac{16}{16 + 0.9836t} = (0.5)^{0.9836} = 0.50572$$

Solving for t, we get t = 15.9. In round numbers this indicates that 16 random numbers would need to be drawn to get the required quota of 8 counties for the sample. This does not differ much from t = 18.27 computed by the more laborious method described previously, in fact, it is rather surprising that this latter method gave such good results for such a small population.