CHAPTER 24 How Large a Sample?

Issues in determining sample size

Sometime in the course of almost every study—preferably early in the planning stage—the researcher must decide how large a sample to take. Deciding the size of sample to take is likely to puzzle and distress you at the beginning of your research career. You have to decide somehow, but there are no simple, obvious guides for the decision.

For example, one of the first studies I worked on was a study of library economics (Fussler and Simon, 1961), which required taking a sample of the books from the library's collections. Sampling was expensive, and we wanted to take a correctly sized sample. But how large should the sample be? The longer we searched the literature, and the more people we asked, the more frustrated we got because there just did not seem to be a clear-cut answer. Eventually we found out that, even though there are some fairly rational ways of fixing the sample size, most sample sizes in most studies are fixed simply (and irrationally) by the amount of money that is available or by the sample size that similar pieces of research have used in the past.

The rational way to choose a sample size is by weighing the benefits you can expect in information against the cost of increasing the sample size. In principle you should continue to increase the sample size until the benefit and cost of an additional sampled unit are equal. [1]

The benefit of additional information is not easy to estimate even in applied research, and it is extraordinarily difficult to estimate in basic research. Therefore, it has been the practice of researchers to set up target goals of the *degree of accuracy* they wish to achieve, or to consider various degrees of accuracy that might be achieved with various sample sizes, and then to balance the degree of accuracy with the cost of achieving that accuracy. The bulk of this chapter is devoted to learning how the sample size is related to accuracy in simple situations.

In complex situations, however, and even in simple situations for beginners, you are likely to feel frustrated by the difficulties of relating accuracy to sample size, in which case you cry out to a supervisor, "Don't give me complicated methods, just give me a rough number based on your greatest experience." My inclination is to reply to you, "Sometimes life is hard and there is no shortcut." On the other hand, perhaps you can get more information than misinformation out of knowing sample sizes that have been used in other studies. Table 24-1 shows the middle (modal), 25th percentile, and 75th percentile scores for—please keep this in mind—*National Opinion Surveys* in the top panel. The bottom panel shows how subgroup analyses affect sample size.

Pretest sample sizes are smaller, of course, perhaps 25-100 observations. Samples in research for Master's and Ph.D. theses are likely to be closer to a pretest than to national samples.

Table 24-1

Most Common Sample Sizes Used for National and Regional Studies By Subject Matter

Subject Matter	**National**			**Regional**		
	Mode	**Q3**	**Q1**	**Mode**	**Q3**	**Q1**
Financial	1000+	—	—	100	400	50
Medical	1000+	1000+	500	1000+	1000+	250
Other Behavior	1000+	—	—	700	1000	300
Attitudes	1000+	1000+	1500	700	1000	400
Laboratory Experiments	—	—	—	100	200	50

Typical Sample Sizes for Studies of Human and Institutional Populations

	People or Households		Institutions	
Subgroup Analyses	**National**	**Special**	**National**	**Special**
Average	1500-2500	500-1000	500-1000	200-500
Many	2500+	1000+	1000+	500+

SOURCE: From *Applied Sampling*, by Seymour Sudman, pp. 86-87. Copyright 1976 by Academic Press, reprinted by permission.

Once again, the sample size ought to depend on the proportions of the sample that have the characteristics you are interested in, the extent to which you want to learn about subgroups as well as the universe as a whole, and of course the purpose of your study, the value of the information, and the cost. Also, keep in mind that the *added* information that you obtain from an additional sample observation tends to be smaller as the sample size gets larger. You must quadruple the sample to halve the error.

Now let us consider some specific cases. The first examples taken up here are from the descriptive type of study, and the latter deal with sample sizes in relationship research.

Some practical examples

Example 24-1

What proportion of the homes in Countryville are tuned into television station WCNT's ten o'clock news program? That is the question your telephone survey aims to answer, and you want to know how many randomly selected homes you must telephone to obtain a sufficiently large sample.

Begin by guessing the likeliest answer, say 30 percent in this case. Do not worry if you are off by 5 per cent or even 10 per cent; and you will probably not be further off than that. Select a first-approximation sample size of perhaps 400; this number is selected from my general experience, but it is just a starting point. Then proceed through the first 400 numbers in the random-number table, marking down a *yes* for numbers 1-3 and *no* for numbers 4-10 (because 3/10 was your estimate of the proportion listening). Then add the number of *yes* and *no*. Carry out perhaps ten sets of such trials, the results of which are in Table 24-2.

Table 24-2

Trial	Number "Yes"	Number "No"	% DIFFERENCE FROM Expected Mean of 30% (120 "Yes")
1	115	285	1.25
2	119	281	0.25
3	116	284	1.00
4	114	286	1.50
5	107	293	3.25
6	116	284	1.00
7	132	268	3.00
8	123	277	0.75
9	121	279	0.25
10	114	286	1.50
Mean			1.37

Based on these ten trials, you can estimate that if you take a sample of 400 and if the "real" viewing level is 30 percent, your average percentage error will be 1.375 percent on either side of 30 percent. That is, with a sample of 400, half the time your error will be greater than 1.375 percent if 3/10 of the universe is listening.

Now you must decide whether the estimated error is small enough for your needs. If you want greater accuracy than a sample of 400 will give you, increase the sample size, using this important rule of thumb: To cut the error in half, you must *quadruple* the sample size. In other words, if you want a sample that will give you an error of only 0.55 percent on the average, you must increase the sample size to 1,600 interviews. Similarly, if you cut the sample size to 100, the average error will be only 2.75 percent (double 1.375 percent) on either side of 30 percent. If you distrust this rule of thumb, run ten or so trials on sample sizes of 100 or 1,600, and see what error you can expect to obtain on the average.

If the "real" viewership is 20 percent or 40 percent, instead of 30 percent, the accuracy you will obtain from a sample size of 400 will not be very different from an "actual" viewership of 30 percent, so do not worry about that too much, as long as you are in the right general vicinity.

Accuracy is *slightly* greater in smaller universes but *only* slightly. For example, a sample of 400 would give *perfect* accuracy if Countryville had only 400 residents. And a sample of 400 will give *slightly* greater accuracy for a town of 800 residents than for a city of 80,000 residents. But, beyond the point at which the sample is a *large fraction* of the total universe, there is no difference in accuracy with increases in the size of universe. This point is very important. For any given level of accuracy, *identical* sample sizes give the same level of accuracy for Podunk (population 8,000) or New York City (population 8 million). The *ratio* of the sample size to the population of Podunk or New York City means nothing at all, even though it intuitively seems to be important.

The size of the sample must depend upon which population or subpopulations you wish to describe. For example, A. Kinsey's sample size would have seemed large, by customary practice, for generalizations about the United States population as a whole. But, as Kinsey explains: "The chief concern of the present study is an understanding of the sexual behavior of *each segment of the population*, and it is only secondarily concerned with generalization for the population as a whole" (Kinsey, et al., 1948, p. 82, italics added). Therefore Kinsey's sample had to include subsamples large enough to obtain the desired accuracy in *each* of these sub-universes. The U.S. Census offers a similar illustration. When the U.S. Bureau of the Census aims to estimate only a total or an average for the United States as a whole—as, for example, in the Current Population Survey estimate of unemployment—a sample of perhaps 50,000 is big enough. But the decennial census aims to make estimates for all the various communities in the country, estimates that require adequate subsamples in each of these sub-universes; such is the justification for the decennial census' sample size of so many millions. Television ratings illustrate both types of purpose. Nielsen ratings, for example, are sold primarily to national network advertisers. These advertisers on national television networks usually sell their goods all across the country and are therefore interested primarily in the total United States viewership for a program, rather than in the viewership in various demographic subgroups. The appropriate calculations for Nielsen sample size will therefore refer to the total United States sample. But other organizations sell rating services to *local* television and radio stations for use in soliciting advertising over the local stations rather than over the network as a whole. Each local sample must then be large enough to provide reasonable accuracy, and, considered as a whole, the samples for the local stations therefore add up to a much larger sample than the Nielsen and other nationwide samples.

The problem may be handled with the following RESAMPLING STATS program. This program represents viewers with random numbers between 1 and 100 and, consistent with our guess that 30% are tuned in, represents viewers with the numbers 1-30. It GENERATES a sample of 400 such numbers, COUNTS the "viewers," then finds how much this sample diverges from the expected number of viewers (30% of 400 = 120). It repeats this procedure 1000 times, and then calculates the average divergence.

```
REPEAT 1000
Do 1000 trials

    GENERATE 400 1,100 a
    Generate 400 numbers between 1 and 100, let 1-30 = viewer

    COUNT a <=30 b
    Count the viewers

    SUBTRACT 120 b c
    How different from expected?

    ABS c d
    Absolute value of difference?

    DIVIDE d 400 e
    Express as a proportion of sample

    SCORE e z
    Keep score of the result

END

MEAN z k
Find the mean divergence
```

Note: The file "tvviewer" on the Resampling Stats software disk contains this set of commands.

It is a simple matter to go back and try a sample size of (say) 1600 rather than 400, and examine the effect on the mean difference.

Example 24-2

This example, like Example 24-1, illustrates the choice of sample size for estimating a summarization statistic. Later examples deal with sample sizes for probability statistics.

Hark back to the pig-ration problems presented earlier, and consider the following set of pig weight-gains recorded for ration A: 31, 34, 29, 26, 32, 35, 38, 34, 31, 29, 32, 30. Assume that our purpose now is to estimate the average weight gain for ration A, so that the feed company can advertise to farmers how much weight gain to expect from ration A. If the universe is made up of pig weight-gains like those we observed, we can simulate the universe with, say, 1 million weight gains of thirty-one pounds, 1 million of thirty-four pounds, and so on for the twelve observed weight gains. Or, more conveniently, as accuracy will not be affected much, we can make up a universe of say, thirty cards for each thirty-one-pound gain, thirty cards for each thirty-four-pound gains and so forth, yielding a deck of 30 x 12 = 360 cards. Then shuffle, and, just for a starting point, try sample sizes of twelve pigs. The means of the samples for twenty such trials are as in Table 24-3.

Now ask yourself whether a sample size of twelve pigs gives you enough accuracy. There is a .5 chance that the mean for the sample will be more than .65 or .92 pound (the two median deviations) or (say) .785 pound (the midpoint of the two medians) from the mean of the universe that generates such samples, which in this situation is 31.75 pounds. Is this close enough? That is up to you to decide in light of the purposes for which you are running the experiment. (The logic of the inference you make here is inevitably murky, and use of the term "real mean" can make it even murkier, as is seen in the discussion in Chapters 20-22 on confidence intervals.)

To see how accuracy is affected by larger samples, try a sample size of forty-eight "pigs" dealt from the same deck. (But, if the sample size were to be much larger than forty-eight, you might need a "universe" greater than 360 cards.) The results of twenty trials are in Table 24-4.

In half the trials with a sample size of forty-eight the difference between the sample mean and the "real" mean of 31.75 will be .36 or .37 pound (the median deviations), smaller than with the values of .65 and .92 for samples of 12 pigs. Again, is this too little accuracy for you? If so, increase the sample size further.

Table 24-3

Trial	Mean	Absolute Devisation of Trial Mean from Actual Mean	Trial	Mean	Absolute Deviation of Trial Mean from Actual Mean
1	31.77	.02	11	32.10	.35
2	32.27	1.52	12	30.67	1.08
3	31.75	.00	13	32.42	.67
4	30.83	.92	14	30.67	1.08
5	30.52	1.23	15	32.25	.50
6	31.60	.15	16	31.60	.15
7	32.46	.71	17	32.33	.58
8	31.10	.65	18	33.08	1.33
9	32.42	.35	19	33.01	1.26
10	30.60	1.15	20	30.60	1.15
Mean					31.75

The attentive reader of this example may have been troubled by this question: How do you know what kind of a distribution of values is contained in the universe before the sample is taken? The answer is that you guess, just as in Example 24-1 you guessed at the mean of the universe. If you guess wrong, you will get either more accuracy or less accuracy than you expected from a given sample size, but the results will not be fatal; if you obtain more accuracy than you wanted, you have wasted some money, and, if you obtain less accuracy, your sample dispersion will tell you so, and you can then augment the sample to boost the accuracy. But an error in guessing will not introduce error into your final results.

Table 24-4

Trial	Mean	Absolute Deviation of Trial Mean from Actual Mean	Trial	Mean	Absolute Deviation of Trial Mean from Actual Mean
1	31.80	.05	11	31.93	.18
2	32.27	.52	12	32.40	.65
3	31.82	.07	13	31.32	.43
4	31.39	.36	14	32.07	.68
5	31.22	.53	15	32.03	.28
6	31.88	.13	16	31.95	.20
7	31.37	.38	17	31.75	.00
8	31.48	.27	18	31.11	.64
9	31.20	.55	19	31.96	.21
10	32.01	.26	20	31.32	.43
Mean					31.75

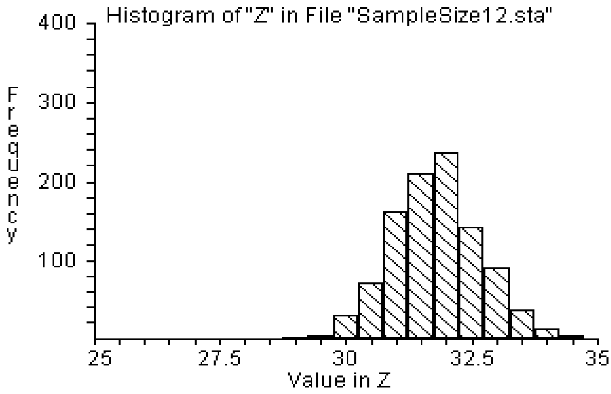
The guess should be based on something, however. One source for guessing is your general knowledge of the likely dispersion; for example, if you were estimating male heights in Rhode Island, you would be able to guess what proportion of observations would fall within 2 inches, 4 inches, 6 inches, and 8 inches, perhaps, of the real value. Or, much better yet, a very small pretest will yield quite satisfactory estimates of the dispersion.

Here is a RESAMPLING STATS program that will let you try different sample sizes, and then take bootstrap samples to determine the range of sampling error. You set the sample size with the DATA command, and the NUMBERS command records the data. Above I noted that we could sample without replacement from a “deck” of thirty “31”’s, thirty “34”’s, etc, as a substitute for creating a universe of a million “31”’s, a million “34”’s, etc. We can achieve the same effect if we replace each card after we sample it; this is equivalent to creating a “deck” of an infinite number of “31”’s, “34”’s, etc. That is what the SAMPLE command does, below. Note that the sample size is determined by the value of the “sampsize” variable, which you set at the beginning. From here on the program takes the MEAN of each sample, keeps SCORE of that result, and produces a HISTOGRAM. The PERCENTILE command will also tell you what values enclose 90% of all sample results, excluding those below the 5th percentile and above the 95th percentile.

Here is a program for a sample size of 12.

```
DATA (12) sampsize

NUMBERS (31 34 29 26 32 35 38 34 32 31 30 29) a

REPEAT 1000

    SAMPLE sampsize a b

    MEAN b c

    SCORE c z

END

HISTOGRAM z

PERCENTILE z (5 95) k

PRINT k
```

```
   Bin Center          Freq               Pct          Cum Pct
```



29.0	2	0.2	0.2
29.5	4	0.4	0.6
30.0	30	3.0	3.6
30.5	71	7.1	10.7
31.0	162	16.2	26.9
31.5	209	20.9	47.8
32.0	237	23.7	71.5
32.5	143	14.3	85.8
33.0	90	9.0	94.8
33.5	37	3.7	98.5
34.0	12	1.2	99.7
34.5	3	0.3	100.0

k = 30.417 33.25

Example 24-3

This is the first example of sample-size estimation for *probability* (testing) statistics, rather than the summarization statistics dealt with above.

Recall the problem of the sex of fruit-fly offspring discussed in Example 15-1. The question now is, how large a sample is needed to determine whether the radiation treatment results in a sex ratio other than a 50-50 male-female split?

The first step is, as usual, difficult but necessary. As the researcher, you must *guess* what the sex ratio will be if the treatment *does* have an effect. Let's say that you use all your general knowledge of genetics and of this treatment and that you guess the sex ratio will be 75 percent males and 25 percent females *if* the treatment alters the ratio from 50-50.

In the random-number table let "01-25" stand for females and "26-00" for males. Take twenty successive pairs of numbers for each trial, and run perhaps fifty trials, as in Table 24-5.

Table 24-5

Trial	Females	Males	Trial	Females	Males	Trial	Females	Males
1	4	16	18	7	13	34	4	16
2	6	14	19	3	17	35	6	14
3	6	14	20	7	13	36	3	17
4	5	15	21	4	16	37	8	12
5	5	15	22	4	16	38	4	16
6	3	17	23	5	15	39	3	17
7	7	13	24	8	12	40	6	14
8	6	14	25	4	16	41	5	15
9	3	17	26	1	19	42	2	18
10	2	18	27	5	15	43	8	12
11	6	14	28	3	17	44	4	16
12	1	19	29	8	12	45	6	14
13	6	14	30	8	12	46	5	15
14	3	17	31	5	15	47	3	17
15	1	19	32	3	17	48	5	15
16	5	15	33	4	16	49	3	17
17	5	15				50	5	15

In Example 15-1 with a sample of twenty flies that contained fourteen or more males, we found only an 8% probability that such an extreme sample would result from a 50-50 universe. Therefore, if we observe such an extreme sample, we rule out a 50-50 universe.

Now Table 24-5 tells us that, if the ratio is *really* 75 to 25, then a sample of twenty will show fourteen or more males forty-two of fifty times (84 percent of the time). If we take a sample of twenty flies and if the ratio is really 75-25, we will make the correct decision by deciding that the split is not 50-50 84 percent of the time.

Perhaps you are not satisfied with reaching the right conclusion only 84 percent of the time. In that case, still assuming that the ratio will really be 75-25 if it is not 50-50, you need to take a sample larger than twenty flies. How much larger? That depends on how much surer you want to be. Follow the same procedure for a sample size of perhaps eighty flies. First work out for a sample of eighty, as was done in Example 15-1 for a sample of twenty, the number of males out of eighty that you would need to find for the odds to be, say, 9 to 1 that the universe is not 50-50; your estimate turns out to be forty-eight males. Then run fifty trials of eighty flies each on the basis of 75-25 probability, and see how often you would not get as many as forty-eight males in the sample. Table 24-6 shows the results we got. No trial was anywhere near as low as forty-eight, which suggests that a sample of eighty is larger than necessary if the split is really 75-25.

Table 24-6

Trial	Females	Males	Trial	Females	Males	Trial	Females	Males
1	21	59	18	13	67	34	21	59
2	22	58	19	19	61	35	17	63
3	13	67	20	17	63	36	22	58
4	15	65	21	17	63	37	19	61
5	22	58	22	18	62	38	21	59
6	21	59	23	26	54	39	21	59
7	13	67	24	20	60	40	21	59
8	24	56	25	16	64	41	21	59
9	16	64	26	22	58	42	18	62
10	21	59	27	16	64	43	19	61
11	20	60	28	21	59	44	17	63
12	19	61	29	22	58	45	13	67
13	21	59	30	21	59	46	16	64
14	17	63	31	22	58	47	21	59
15	22	68	32	19	61	48	16	64
16	22	68	33	10	70	49	17	63
17	17	63				50	21	59

Table 24-7

Trial	Females	Males	Trial	Females	Males	Trial	Females	Males

1	35	45	18	32	48	34	35	45
2	36	44	19	28	52	35	36	44
3	35	45	20	32	48	36	29	51
4	35	45	21	33	47	37	36	44
5	36	44	22	37	43	38	36	44
6	36	44	23	36	44	39	31	49
7	36	44	24	31	49	40	29	51
8	34	46	25	27	53	41	30	50
9	34	46	26	30	50	42	35	45
10	29	51	27	31	49	43	32	48
11	29	51	28	33	47	44	30	50
12	32	48	29	37	43	45	37	43
13	29	51	30	30	50	46	31	49
14	31	49	31	31	49	47	36	44
15	28	52	32	32	48	48	34	64
16	33	47	33	34	46	49	29	51
17	36	44				50	37	43

It is obvious that, if the split you guess at is 60 to 40 rather than 75 to 25, you will need a bigger sample to obtain the "correct" result with the same probability. For example, run some eighty-fly random-number trials with 1-40 representing males and 51-100 representing females. Table 24-7 shows that only twenty-four of fifty (48 percent) of the trials reach the necessary cut-off at which one would judge that a sample of eighty really does not come from a universe that is split 50-50; therefore, a sample of eighty is not big enough if the split is 60-40.

To review the main principles of this example: First, the closer together the two possible universes from which you think the sample might have come (50-50 and 60-40 are closer together than are 50-50 and 75-25), the larger the sample needed to distinguish between them. Second, the surer you want to be that you reach the right decision based upon the sample evidence, the larger the sample you need.

The problem may be handled with the following RESAMPLING STATS program. We construct a benchmark universe that is 60-40 male-female, and take samples of size 80, observing whether the numbers of males and females differs enough in these resamples to rule out a 50-50 universe. Recall that we need at least 48 males to say that the proportion of males is *not* 50%.

```
REPEAT 1000
```

Do 1000 trials

```
    GENERATE 80 1,10 a
    Generate 80 “flies,” each represented by a number between 1 and 10
    where <= 6 is a male

    COUNT a <=6 b
    Count the males

    SCORE b z
    Keep score

END

COUNT z >=48 k
How many of the trials produced more than 48 males?

DIVIDE k 1000 kk
Convert to a proportion

PRINT kk
```

If the result “kk” is close to 1, we then know that samples of size 80 will almost always produce samples with enough males to avoid misleading us into thinking that they could have come from a universe in which males and females are split 50-50.

Example 24-3

Referring back to Example 15-3, on the cable-television poll, how large a sample *should* you have taken? Pretend that the data have not yet been collected. You need *some* estimate of how the results will turn out before you can select a sample size. But you have not the foggiest idea how the results will turn out. Therefore, go out and take a very small sample, maybe ten people, to give you some idea of whether people will split quite evenly or unevenly. Seven of your ten initial interviews say they are for CATV. How large a sample do you now need to provide an answer of which you can be fairly sure?

Using the techniques of the previous chapter, we estimate roughly that from a sample of fifty people at least thirty-two would have to vote the same way for you to believe that the odds are at least 19 to 1 that the sample does not misrepresent the universe, that is, that the sample does not show a majority different from that of the whole universe if you polled everyone. This estimate is derived from the resampling experiment described in example 15-3. The table shows that if half the people (or more) are against cable television, only one in twenty times will thirty-two (or more) people of a sample of fifty say that they are for cable television; that is, only one of twenty trials with a 50-50 universe will produce as many as thirty-two *yeses* if a majority of the population is against it.

Therefore, designate numbers 1-30 as *no* and 31-00 as *yes* in the random-number table (that is, 70 percent, as in your estimate based on your presample of ten), work through a trial sample size of fifty, and count the number of *yeses*. Run through perhaps ten or fifteen trials, and reckon how often the observed number of *yeses* exceeds thirty-two, the number you must exceed for a result you can rely on. In Table 24-8 we see that a sample of fifty respondents, from a universe split 70-30, will show that many *yeses* a preponderant proportion of the time—in fact, in fifteen of fifteen experiments; therefore, the sample size of fifty is large enough if the split is "really" 70-30.

Table 24-8

Trial	No	Yes	Trial	No	Yes
1	13	37	9	15	35
2	14	36	10	9	41
3	18	32	11	15	35
4	10	40	12	15	35
5	13	37	13	9	41
6	15	35	14	16	34
7	14	36	15	17	33

The following RESAMPLING STATS program takes samples of size 50 from a universe that is 70% "yes." It then observes how often such samples produce more than 32 "yeses"—the number we must get if we are to be sure that the sample is not from a 50/50 universe.

```
REPEAT 1000
Do 1000 trials

    GENERATE 50 1,10 a
    Generate 50 numbers between 1 and 10, let 1-7 = yes.

    COUNT a <=7 b
    Count the "yeses"

    SCORE b z
    Keep score of the result

END

COUNT z >=32 k
Count how often the sample result >= our 32 cutoff (recall that samples
with 32 or fewer "yeses" cannot be ruled out of a 50/50 universe)

DIVIDE k 1000 kk
Convert to a proportion
```

If "kk" is close to 1, we can be confident that this sample will be large enough to avoid a result that we might mistakenly think comes from a 50/50 universe (provided that the real universe is 70% favorable).

Example 24-4

How large a sample is needed to determine whether there is any difference between the two pig rations in Example 15-7? The first step is to guess the results of the tests. You estimate that the average for ration A will be a weight gain of thirty-two pounds. You further guess that twelve pigs on ration A might gain thirty-six, thirty-five, thirty-four, thirty-three, thirty-three, thirty-two, thirty-two, thirty-one, thirty-one, thirty, twenty-nine, and twenty-eight pounds. This set of guesses has an equal number of pigs above and below the average and more pigs close to the average than farther away. That is, there are more pigs at 33 and 31 pounds than at 36 and 28 pounds. This would seem to be a reasonable distribution of pigs around an average of 32 pounds. In similar fashion, you guess an average weight gain of 28 pounds for ration B and a distribution of 32, 31, 30, 29, 29, 28, 28, 27, 27, 26, 25, and 24 pounds.

Let us review the basic strategy. We want to find a sample size large enough so that a large proportion of the time it will reveal a difference between groups big enough to be accepted as not attributable to chance. First, then, we need to find out how big the difference must be to be accepted as evidence that the difference is not attributable to chance. We do so from trials with samples that size from the benchmark universe. We state that a difference larger than the benchmark universe will usually produce is not attributable to chance.

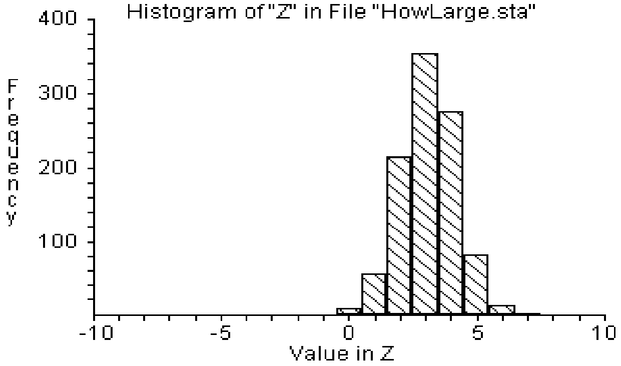
In this case, let us try samples of 12 pigs on each ration. First we draw two samples from a combined benchmark universe made up of the results that we have guessed will come from ration A and ration B. (The procedure is the same as was followed in Example 15-7.) We find that in 19 out of 20 trials the difference between the two observed groups of 12 pigs was 3 pounds or less. Now we investigate how often samples of 12 pigs, drawn from the *separate* universes, will show a mean difference as large as 3 pounds. We do so by making up a deck of 25 or 50 cards for *each* of the 12 hypothesized A's and each of the 12 B's, with the ration name and the weight gain written on it—that is, a deck of, say, 300 cards for each ration. Then from each deck we draw a set of 12 cards at random, record the group averages, and find the difference.

Here is the same work done with more runs on the computer:

```
NUMBERS (31 34 29 26 32 35 38 34 32 31 30 29) a

NUMBERS (32 32 31 30 29 29 29 28 28 26 26 24) b

REPEAT 1000

     SAMPLE 12 a aa

     MEAN aa aaa

     SAMPLE 12 b bb

     MEAN bb bbb

     SUBTRACT aaa bbb c

     SCORE c z

END

HISTOGRAM z
```

Difference in mean weights between resamples

Therefore, two samples of twelve pigs each are clearly large en fficient



if the universes are really like those we guessed at. If, on the other hand, the differences in the guessed universes had been smaller, then twelve-pig groups would have seemed too small and we would then have had to try out larger sample sizes, say forty-eight pigs in each group and perhaps 200 pigs in each group if forty-eight were not enough. And so on until the sample size is large enough to promise the accuracy we want. (In that case, the decks would also have to be much larger, of course.)

If we had guessed different universes for the two rations, then the sample sizes required would have been larger or smaller. If we had guessed the averages for the two samples to be closer together, then we would have needed larger samples. Also, if we had guessed the weight gains *within* each universe to be less spread out, the samples could have been smaller and vice versa.

The following RESAMPLING STATS program first records the data from the two samples, and then draws from decks of infinite size by sampling with replacement from the original samples.

```
DATA (36 35 34 33 33 32 32 31 31 30 29 28) a

DATA (32 31 30 29 29 28 28 27 27 26 25 24) b

REPEAT 1000

     SAMPLE 12 a aa
     Draw a sample of 12 from ration a with replacement (this is like drawing from a large deck made up of many replicates of the elements in a)

     SAMPLE 12 b bb
     Same for b

     MEAN aa aaa
     Find the averages of the resamples

     MEAN bb bbb

     SUBTRACT aaa bbb c
     Find the difference

     SCORE c z

END

COUNT z >=3 k
How often did the difference exceed the cutoff point for our significance test of 3 pounds?

DIVIDE k 1000 kk

PRINT kk
```

If kk is close to zero, we know that the sample size is large enough that samples drawn from the universes we have hypothesized will not mislead us into thinking that they could come from the same universe.

Step-wise sample-size determination

Often it is wisest to determine the sample size as you go along, rather than fixing it firmly in advance. In sequential sampling, you *continue* sampling until the split is sufficiently even to make you believe you have a reliable answer.

Related techniques work in a series of jumps from sample size to sample size. Step-wise sampling makes it less likely that you will take a sample that is much larger than necessary. For example, in the cable-television case, if you took a sample of perhaps fifty you could see whether the split was as wide as 32-18, which you figure you need for 9 to 1 odds that your answer is right. If the split were not that wide, you would sample another fifty, another 100, or however large a sample you needed until you reached a split wide enough to satisfy you that your answer was reliable and that you really knew which way the entire universe would vote.

Step-wise sampling is not always practical, however, and the cable-television telephone-survey example is unusually favorable for its use. One major pitfall is that the *early* responses to a mail survey, for example, do *not* provide a random sample of the whole, and therefore it is a mistake simply to look at the early returns when the split is not wide enough to justify a verdict. If you have listened to early radio or television reports of election returns, you know how misleading the reports from the first precincts can be if we regard them as a fair sample of the whole.[2]

Stratified sampling is another device that helps reduce the sample size required, by balancing the amounts of information you obtain in the various strata. (Cluster sampling does not reduce the sample size. Rather, it aims to reduce the cost of obtaining a sample that will produce a given level of accuracy.)

Summary

Sample sizes are too often determined on the basis of convention or of the available budget. A more rational method of choosing the size of the sample is by balancing the diminution of error expected with a larger sample, and its value, against the cost of increasing the sample size. The relationship of various sample sizes to various degrees of accuracy can be estimated with resampling methods, which are illustrated here.

Endnotes

1. R. Schlaifer (1961) attacks the sample-size problem in the wider context of decision making, costs, and benefits. The statistically knowledgeable reader can find an excellent discussion of sample size in M. Hansen, et al. A. Mace gives many examples of the appropriate calculation in an engineering framework.

2. See J. Lorie and H. Roberts (pp. 155-157) for more discussion of the limitations of sequential sampling. And M. Hansen, et al., warn against the danger of increasing the sample size in this fashion:

> The investigator examines the returns from an initial sample to determine whether they appear acceptable to the investigator; if they do, he uses the results as they are; if they do not, he discards the sample results [or keeps the old sample] and draws a new sample, perhaps by a different method, in the hope that he will obtain a result more nearly like the one he expected. Such an approach can be utilized to obtain almost any results desired, or can "prove" any point even when unbiased or consistent methods of selecting the sample and making the individual estimates are used if the initial results are subject to relatively large sampling errors. (Hansen, et al., 1953: 78)