Chapter Two

RANDOM SAMPLING AND THE NORMAL DISTRIBUTION*

This chapter introduces the following basic statistical concepts: population, sample, parameter, statistic, random sample and normal distribution. First, a brief explanation of the difference between a population as a whole and a sample drawn from this population will be presented. Parameters will be associated with populations, statistics, with samples. Then, with a deck of playing cards representing the population to be examined, some problems researchers face when they try to draw samples from populations will be demonstrated. In the process, two techniques for obtaining a random sample will be discussed. Finally, data generated by workshop participants at the Seminar on Statistics for Language Studies (Ustronie, Poland, 1988) will be compared.

POPULATIONS VS. SAMPLES

It is normally impossible to examine a whole population. Imagine, for example, trying to conduct an experiment on all university students of English as a Foreign Language in Poland. And this is by no means an unusually large sample in terms of those typically under study in theoretical and applied linguistics (quit the contrary). Thus, we have to rely on samples. Samples - by their very nature - present an incomplete view of the population from which they are drawn. Suppose we draw a sample from a given population and calculate an average for some variable in

---

* Douglas W. Coleman, University of Łódź.

the sample - like an average test score for a sample of Polish university-level students of English as a Foreign Language. It will almost certainly be different from the overall population average. This actual average for the population as a whole (which we will probably never know precisely) is referred to as a parameter. The sample average, which is an estimate of this parameter, is a statistic. The quality of this estimate, to a large degree, depends on how representative the sample is of the population as a whole. At this point, then, let's turn to the problem of how to obtain a representative sample.

RANDOM SAMPLES

Frequently, the best way to obtain a representative sample of a population is to draw a random sample from it. For the sake of illustration, let's suppose the population under study is an ordinary deck of playing cards (54 cards - two jokers included). Suppose we want a random sample of 13 cards. (Normally, of course, a sample is not so large a proportion of the population, but then rarely does a population have only 54 members). We can just take the first 13 cards in the deck 'at random', can't we? Let's try. Here is what we get:

AS 2S 3S 4S 5S 6S 7S 8S 9S 10S JS QS KS

(Here, 'A' = 'ace', 'J' = 'jack', 'Q' = 'queen', 'K' = 'king', and 'S' = 'of spades'.)

Obviously, this does not produce a very representative sample. In fact, although we said we were taking the first 13 cards from the deck 'at random', we did not get a random sample. Why not? A random sample is one in which every member of the population has an equal chance of being selected. Because of the way the cards are ordered in a new deck, taking the first 13 cards 'at random' is not random at all. The result is the same 13 cards every time (the spades); each card thus does not have an equal chance of being selected.

One commonly-used way to get around this is to spread the selection of the sample throughout the whole population, for example, by taking every fourth card. Let's try. Here is what we get this time:

AS 5S 9S KS 4D 8D QD 3C 7C JC 2H 6H 10H

(Now, 'D' = 'of diamonds', 'C' = 'of clubs', and 'H' = 'of hearts'.) This sample looks much more representative.

Usually this procedure works quite well, but not always. Let's try taking every fourth card from another deck. This time we get the following:

AS 2S 3S 4S 5S 6S 7S 8S 9S 10S JS QS KS

This time we have a problem. Now you might want to say we are playing with a 'stacked deck', and it is true. The deck is biased. But the same sort of thing can happen in actual research situations.

For example, let's suppose we're conducting a study of Polish school children, learning English as a Foreign Language. We might ask the school teachers in a given city to provide lists of their students so that we can draw a sample. Suppose we know that the average class size is about 20. If we plan to select every tenth student, we know we can tell the teachers that we will be examining one or two students from each class. Let's suppose that some of the teachers suspect our motives. (Perhaps, despite what we tell them, some of the teachers think the results will be used to evaluate the quality of their instruction). Thus, they might, for instance, order their lists of students from best to worst, hoping that we will simply select one or two students 'at random' from the top of the list. You can see that if we just string these lists together, they will have a cyclical pattern very similar to that in a deck of cards. In this case, it will be very likely that the result will be an unrepresentative sample. Why? If the variation in the class size is small (i.e., if all the classes are very close to the average size of 20), then we will tend to get all our subjects from (a) near the top end of each class and (b) near the middle (median) for each class. This will bias any estimates we make, such as our sample average, significantly upwards. (We could solve this particular problem by reordering the lists alphabetically before drawing our sample; but there is a more general approach).

USING RANDOM NUMBER TABLES

Another commonly-used technique for obtaining a random sample is to use a list of random numbers to select items from the population. One way to do this is to use a table of random numbers from the back of a statistics handbook, such as Shavelson (1981). To select a random sample of 13 items from our population of 54 playing cards, one possibility would be to write down the first 13 unique numbers in such a table which fall between 1 and .54. (i.e., we would reject numbers like 78, 55, or 90). Following this procedure, we might get the following numbers:

22 17 23 35 02 51 09 43 06 24 03 47 19

Sorted for easier use (and with left-hand zeros removed), here is the same list:

2 3 6 9 17 19 22 23 24 35 43 47 51

Thus, we would take the second, third, sixth, ninth cards, and so on, from the deck.

Another slightly more complicated procedure uses all the numbers from the table. Assume the random number table consists of a list of 2-digit numbers, all of which therefore fall between 00 (zero) and 99. We want only numbers between 1 and 54. We can add one to each number we find, then multiply by the fraction .54 (54/100). If the result contains a fraction, we will round it off to the nearest whole number. Thus, suppose the first value in the table is 78. The calculation is

$$(78 + 1) \times .54 = 42.66$$

- which rounds up to 43. So, our sample will include the forty--third item. Suppose the next number in the table is 22. This time,

$$(22 + 1) \times .54 = 12.42$$

- which rounds down to 12, meaning our sample will also include the twelfth card. We can continue in this way until we have selected 13 different cards from the deck.

For the sake of illustration, let's take the set of numbers obtained the first way and see what kind of a sample we get from that first deck of cards. Here it is:

item	2	3	6	9	17	19	22	23	24	35	43	47	51
card	2S	3S	6S	9S	4D	6D	9D	10D	JD	5C	10H	6H	2H

How representative a sample is this? Of course, no sample is perfectly representative. It can't be. But it does look like we have managed to avoid obvious bias in the sample. The median rank in this sample is 6. The population median is (given as a whole number) 7. (This assumes we count an ace as 1, a jack as 10, a queen as 11, a king as 12, and a joker as 13.) The sample median is a fair estimate of the population median. Let's treat the suits as nominal data. How well does the sample represent the proportions of the suits (spades, diamonds, clubs, hearts) in the population as a whole? If we count the jokers as a separate 'suit', then we have these proportions in the population (the whole deck):

spades	13/54 =	.24
diamonds	13/54 =	.24
clubs	13/54 =	.24
hearts	13/54 =	.24
jokers	2/54 =	.04
		1.00 total

since proportion equals frequency divided by the population size (figures are rounded to two decimal places). Our sample does not seem to have given a very accurate picture of the population in this regard. Here is what we got for the sample:

spades	4/13 =	.31
diamonds	5/13 =	.38
clubs	1/13 =	.08
hearts	3/13 =	.23
jokers	0/13 =	.00
		1.00 total

(Here, proportion equals frequency divided by the sample size; again, figures are rounded off.) On the other hand, we did do much better than in drawing the two earlier samples in which we observed only spades.

Now, let's try the same procedure on the second deck of cards we used earlier. Just for the sake of comparison, we will even use the same set of random numbers. Here is our sample:

item	2	3	6	9	17	19	22	23	24	35	43	47	51
card	AS	AD	2S	3H	5H	5D	6C	6D	6S	9D	JD	QD	KD

Again, the sample median is again 6, and again, the proportions of suits in the sample do not match those in the population as a whole very closely:

spades	3/13 =	.23
diamonds	7/13 =	.54
clubs	1/13 =	.08
hearts	2/13 =	.15
jokers	0/13 =	.00
		1.00 total

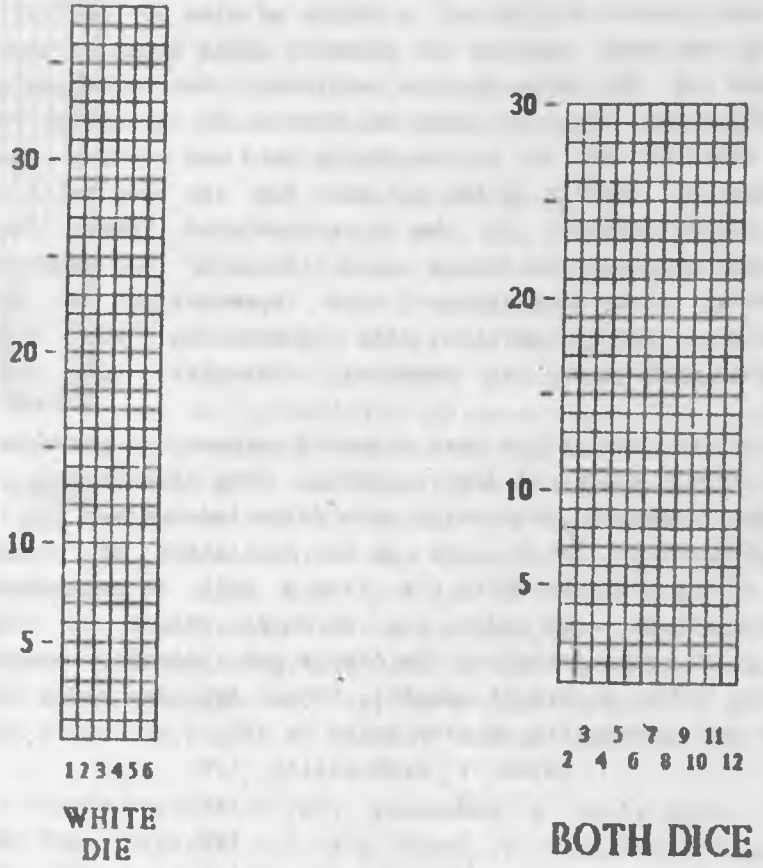

Fig. 1. Tally sheet

THE EFFECT OF SAMPLE SIZE

In a workshop at the Seminar on Statistics for Language Studies (Ustronie, Poland, 1988), participants examined the effect of sample size on the representativeness of the sample. Actually two populations were examined simultaneously. The first was the population of values to be obtained on a roll of two dice. The second was the population of values to be obtained on a roll of just one die. The participants worked in groups; each group had one white die, one red die, and a tally sheet (see Figure 1). Each group was instructed to roll its dice 50 times - equivalent to 50 experimental trials, or a sample of size 50 (actually, two of the groups took samples of about 75 rolls each). On each trial the value of the white die was recorded in the left-hand grid of the tally sheet by blackening one square in the column for that value, starting at the bottom of the grid and working upwards. The value of the sum of the two dice for the same trial was recorded in the same way in the right-hand grid. Thus, the end result of tallying the values was a histogram (see also Chapter One, above) with the horizontal axis representing the value of the die/dice and the vertical axis representing the frequency with which each value was observed. (Examples will be seen below).

These two populations were selected as examples possessing two very different kinds of distributions. (The distribution represents the frequency with which each value is observed in the population). First, let's look at the population of values of rolls of one die, the white die (let's call it population WD). In population WD, each value has an equal chance of occurring on any given roll (assuming the die is not loaded, weighted or shaped to favor a certain result). There are six sides on the die, so the probability of each value is 1/6:

value	probability
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
	1 total

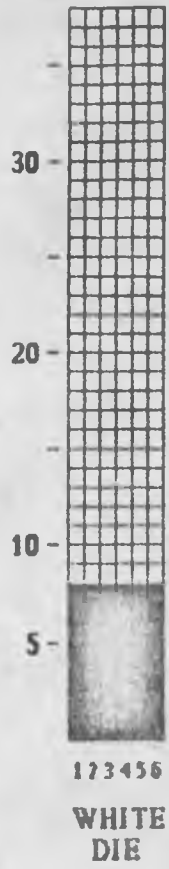

Fig. 2. Population WD

Notice that the total of the probabilities always adds up to one. Given the probabilities here, the histogram for population WD is (theoretically, at least) 'flat'; i.e., we might expect the histogram for a sample from WD of 48 rolls to look like Figure 2. Actual samples, however, are presented below. Very often, nominal level data (e.g., sex, native language, learning style, dialect, etc.) may possess a distribution very similar to that of population WD. Ordinal level data may also have a similar distribution (e.g., Level in school, number of foreign languages studied, and so on). In either case, it may not be a 'flat' distribution, but much of what follows will still apply.

The second population, that of the values of the sum of both dice, let's call population BD. Here, the distribution is quite different. Values range from 2 to 12 (rather than from 1 to 6, as in population WD). More importantly, values do not all possess the same probability of occurring. For example, there is only one way for the value 2 to occur - with the white die having a value of 1 and the red die having the same value. But there are six possible combinations of the red and white dice that add up to the value 7. Figure 3 shows (a) all the possible combinations of two dice and (b) how the probability of each value can be counted up. Thus, from a sample of size 36, we might expect a histogram like that shown in Figure 3. Again, actual samples are shown below.

This second population (BD) possesses a distribution very much like that referred to as a normal distribution, shown in Figure 4. The normal distribution has three major characteristics. First, the three common measures of central tendency - the mean, median, and mode (explained in the first and fourth chap-

Combined value	Red	White	Probability
2	1	1	(1/6 x 1/6) x 1 = 1/36
3	1	2	(1/6 x 1/6) x 2 = 2/36
	2	1	
4	1	3	(1/6 x 1/6) x 3 = 3/36
	2	2	
	3	1	
5	1	4	(1/6 x 1/6) x 4 = 4/36
	2	3	
	3	2	
	4	1	
6	1	5	(1/6 x 1/6) x 5 = 5/36
	2	4	
	3	3	
	4	2	
	5	1	
7	1	6	(1/6 x 1/6) x 6 = 6/36
	2	5	
	3	4	
	4	3	
	5	2	
	6	1	
8	2	6	(1/6 x 1/6) x 5 = 5/36
	3	5	
	4	4	
	5	3	
	6	2	
9	3	6	(1/6 x 1/6) x 4 = 4/36
	4	5	
	5	4	
	6	3	
10	4	6	(1/6 x 1/6) x 3 = 3/36
	5	5	
	6	4	
11	5	6	(1/6 x 1/6) x 2 = 2/36
	6	5	
12	6	6	(1/6 x 1/6) x 1 = 1/36
			total = 1 = 36/36

Fig. 3

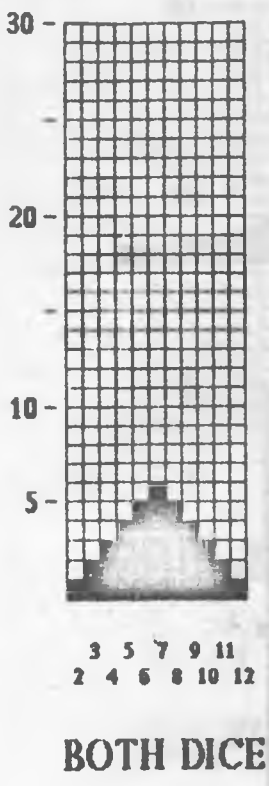

Fig. 4. Population BD

ters of this volume) - are all equal. Second, the distribution is symmetrical and 'bell-shaped'. Finally, there are no zero frequencies; the right and left ends of the curve (referred to as the 'tails' of the distribution) never quite reach the horizontal axis. The second and third of these characteristics are obvious in the histogram for a theoretically 'ideal' sample from population BD (Figure 4). The first characteristics can also be confirmed, but this will be left as an exercise for the reader. (It is necessary to imagine that all the values, 2 through 12, represent equal interval level data, not merely ordinal data, to calculate a mean.) The data for calculating the mean, median and mode can be obtained from the histogram.

The workshop participants who rolled the dice and tallied the results worked in five groups. Thus, five samples were obtained for each of the two populations (WD and BD). Two of these samples for each population are shown in Figures 6 (WD) and 7 (BD). Notice that in neither case is the histogram for a sample from WD 'flat'. And in neither case does the histogram for a sample from BD have a shape like that of a normal distribution. The variation in the shape of the histograms is due to chance. Each roll of the dice is independent: the results of one roll have no effect on the next roll. The probabilities shown in Figure 3, for example (for BD), hold for any given roll of the dice. This means that they will probably hold over the long run. But there is no guarantee they will hold in a sample of, say, 50 rolls, or even in a sample of, say, 1000 rolls. However, it does seem reasonable to expect larger samples to have distributions more closely resembling the 'theoretical' distribution of the population

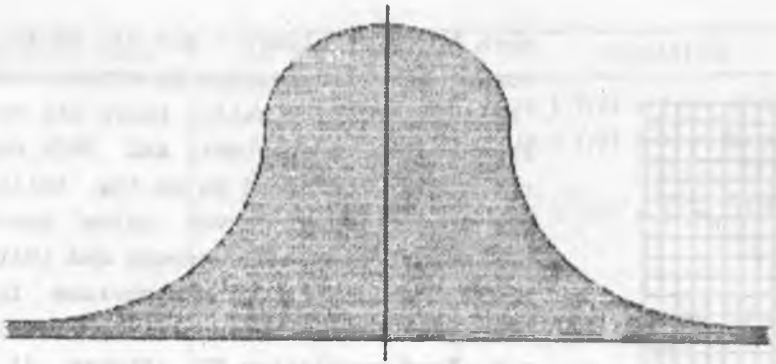

Fig. 5. Normal distribution

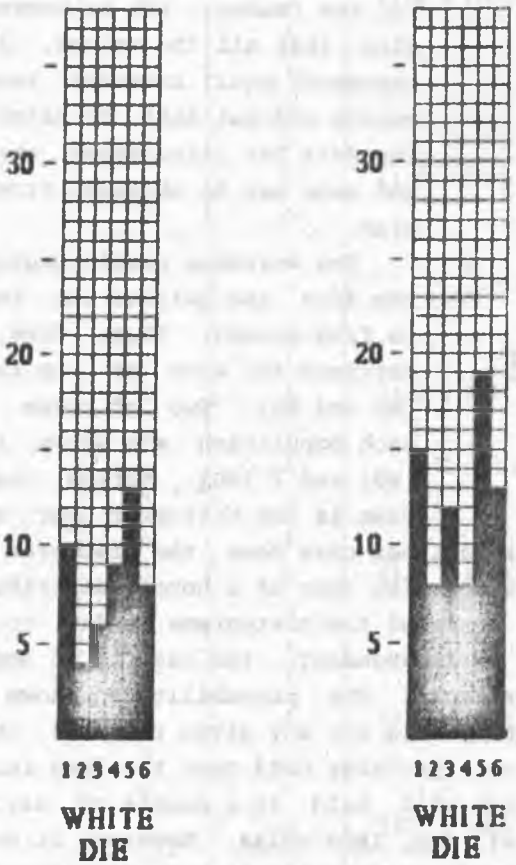

Fig. 6. Tallies for two WD samples

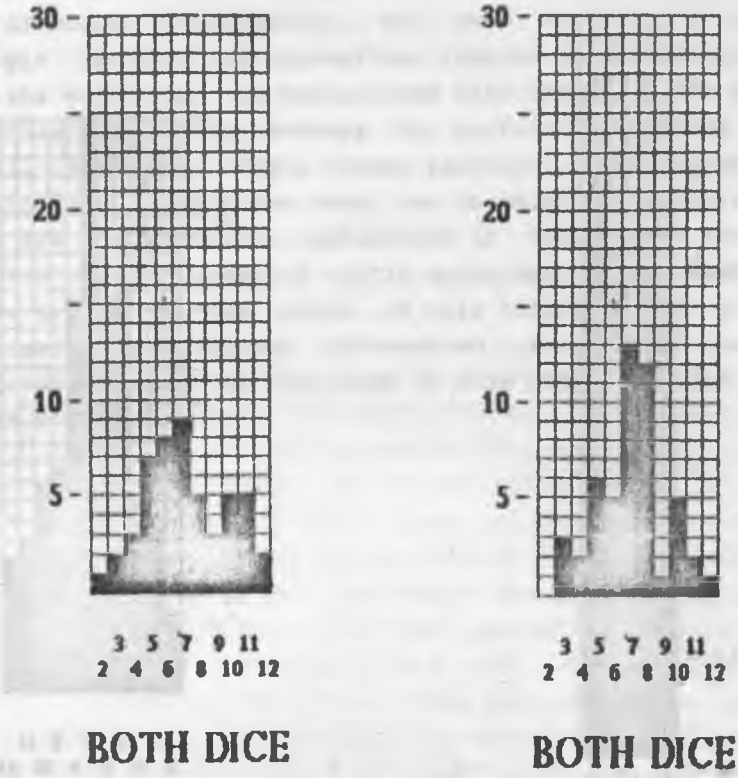

Fig. 7. Tallies for two BD samples

Let's see if this is the case with the samples obtained by the workshop participants. Figure 8 shows the combined tallies for all five groups (302 rolls of the dice). Even with samples of size 302, we still do not have a totally 'flat' distribution for WD or a perfectly symmetrical distribution for BD. However, there does seem to be some improvement - the larger samples do bear a greater resemblance to the theoretical distributions.

This is not necessarily always the case. We might 'get lucky' and draw a sample of size 50 that bears a stronger resemblance to the overall population under study than with another sample of size 100. The larger the sample, the higher the probability, however, that it will be a representative sample.

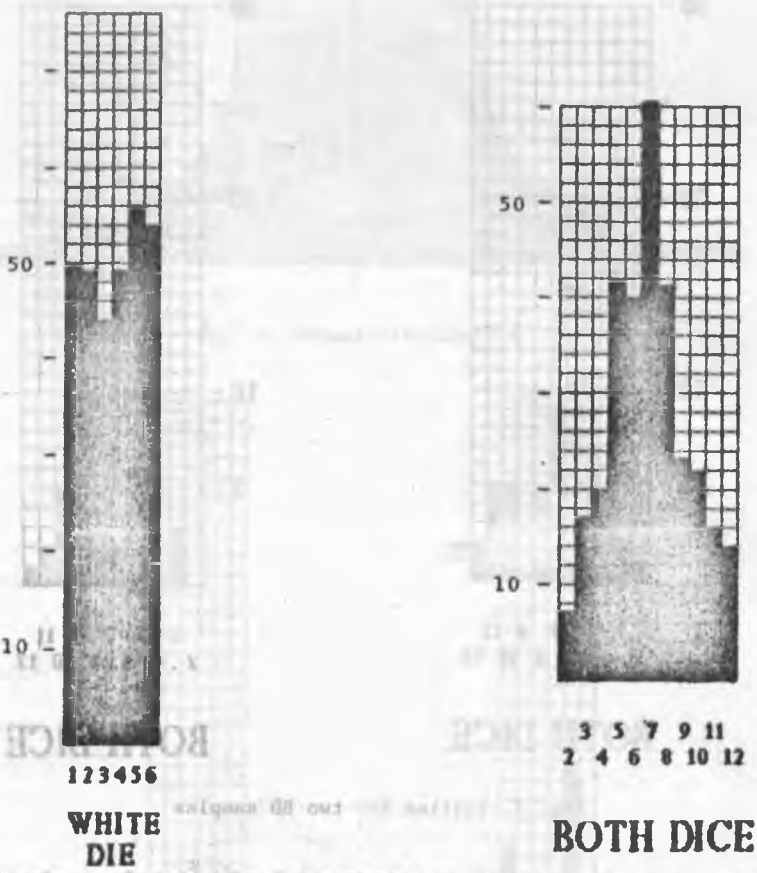

Fig. 8. Combined tallies

This does not mean that a larger sample is always 'better'. In real research situations, the size of the sample may be limited by practical considerations. For example, only a certain number of subjects may be available for testing. The form of the test (or questionnaire, etc.) required by a particular research design may be too time-consuming for use with very many subjects. Time is also a factor when selecting a textual corpus for study. Cost can be a major limiting factor as well. Also, some studies involve populations with clearly identifiable subgroups, each with distinct sets of characteristics (e.g., groups of subjects

separable by regional or social dialect, texts separable by the types of discourse they contain). For such studies, a simple random sample is often not appropriate (unless the researcher can show that the subgroups are homogeneous with regard to the particular variable under study, perhaps by performing a pilot study beforehand). Instead of simple random sampling, a stratified sampling technique is sometimes used, one in which subgroups of the population are proportionally represented in the sample (but individuals are randomly selected within subgroups). Many such considerations are beyond the scope of this chapter, and readers are encouraged to seek further information from other sources (the references section at the back of this book includes some good places to begin).