IV. SAMPLING FROM A POPULATION

The Necessity of Sampling:

As we mentioned in the discussion of populations, it is often not feasible to examine each member of a population. This can be because of the size of the population or because some of its members are difficult to reach. However, it is feasible to examine the individuals in a smaller subset of the population. A subset of the population is called a sample. Inferential statistics concerns methods of extending conclusions derived from the sample to the whole population.

Example 4.1:

Suppose we are interested in the median travel distance from home to school for students at this university. It would be very hard to determine the value of the median exactly. However, we could stand outside the student center at noon on Wednesday and ask 50 passers-by to tell us their travel distances. We might then take the median of these 50 responses to be an estimate of the population median. This is an example of a statistical inference, albeit a flawed and incomplete one.

The reliability of a statistical inference should be quantifiable in some way. One way to quantify the reliability of an inference procedure is to choose the sample in such a way that a probability can be assigned to each possible result. Then a probability can be assigned to the event that the difference between the sample median and the population median is within a given error tolerance. Suppose for the sake of argument that the probability of achieving the given degree of accuracy is 0.95. Imagine repeating the experiment of choosing a sample many times. We could say that, in the long run, for 95% of the repeated sampling experiments the sample median would meet our standard of accuracy. For any particular sample we cannot say whether the sample median is sufficiently accurate or not. We can only say that we used a sampling method that has a high probability of success.

Discussion Topic: What systematic errors might be involved in interviewing the first 50 passers-by outside the student center at noon on Wednesday to estimate the median travel distance for all students? Do evening students have the same chance of being included in the sample as day students?

For large or partly inaccessible populations, such as the population of U.S. registered voters, statisticians have devised ingenious and intricate methods of sampling. We shall consider only the most basic method – simple random sampling.

Sampling with Replacement:

Let the number of individuals in a population be N. An ordered sample with replacement is a sequence of n individuals from the population. It is called a sample with replacement because repetitions are allowed in the sequence of individuals chosen. The length n of the sequence is the sample size. The individuals in the sample are chosen sequentially and at each stage each individual has probability of $1/N$ of being chosen.

The sample space for this random experiment consists of all ordered sequences of length n of individuals from the population. The number of outcomes of the experiment is N^n. Since the outcomes are equally likely, each outcome has probability $1/N^n$. If $n = 1$, the sample space is the same as the population.

Let x denote a population variable and let X_1 be the value of x for the first individual in the sample, X_2 the value of x for the second individual, and so on. We denote these by upper-case letters because they are random variables. The distribution of each of them is the same as the distribution of values of x in the population. That is, if x' is one of the values of x, then $P[X_j = x']$ is equal to the relative frequency of occurrence of the value x' in the population. Not only do the X_j s all have the same distribution, they are independent as well. In general, if a random experiment gives rise to a random variable X and that experiment is repeated n times independently, then the resulting sequence $X_1, X_2, \cdots, X_n$ is called a random sample from the distribution of X.

Sampling without Replacement:

An ordered sample without replacement is a sequence *without repetitions* of individuals from the population. If there are N individuals in the population and the sample size is n, then there are $N(N-1)\cdots(N-n+1)$ possible ordered samples without replacement. Suppose that the first individual of the sample is chosen so that all of the individuals are equally likely to be chosen. Having chosen the first individual in the sample, suppose that all of the $N-1$ remaining individuals are equally likely to be chosen next. In general, suppose that once the first m individuals are chosen, each of the remaining $N-m$ are equally likely to be chosen next. We have just described a compound random experiment with equally likely outcomes. Each outcome is an ordered sample, without replacement, of size n and each possible outcome has probability

$$1/N(N-1)\cdots(N-n+1).$$

Under certain circumstances the distinction between sampling with and without replacement is not important. If the population size N is a great deal larger than the sample size n, there is effectively no difference between them. The reason is that even if sampling is done with replacement, so that repetitions could occur, the probability of any repetitions actually occurring is very small.

Example 4.2:

If $N = 1000$ and $n = 10$ and sampling is done with replacement, the probability of one or more repetitions in the sample is 0.044. If $N = 10000$ and $n = 100$, the probability of one or more repetitions is 0.39.

This example shows that more than just the ratio *n/N* is involved. It can be shown that in sampling with replacement, if $n^2 \le N$ then the probability of one or more repetitions is less than $n^2 / 2N$. Therefore, if this number is very small one can safely ignore the distinction. Depending on the kind of statistical inference one has in mind, it may be possible to relax this requirement even more.

Activity: Revisit the birthday problem. $N = 365$. Find the probability of repetitions for several sample sizes *n*.

Usually, when sampling without replacement, the order that the individuals are chosen in is of no importance. All that matters is the resulting subset of *n* individuals from the population. A subset of *n* individuals from the population, chosen as described above, is what is usually meant by the expression "simple random sample of size *n*". The number of subsets of size *n* in a population of size *N* is

$$C_{N,n} = \frac{N!}{n!(N-n)!}.$$

Thus, in the experiment of choosing a simple random sample of size n, the outcomes are equally likely and each outcome has probability

$$\frac{1}{C_{N,n}}.$$

Random Number Generators:

Almost all calculators and spreadsheet programs are capable of simulating random samples. The output is called a pseudo-random sequence because once the sequence is started there is really no element of chance involved. The sequence is supposed to be a randomly selected sample from the set of numbers between 0 and 1, rounded but expressed to a high precision.

Exercise: Press the random number key of your calculator 50 times and record the random numbers generated. Draw a histogram of the resulting data with class intervals 0 to .2, .2 to .4, etc.

Suppose you want to generate a random sample, with replacement, of length n from a population with N individuals. Number the individuals from 1 to N in any way that is convenient. Then generate *n* pseudo-random numbers. If a random number falls between $\frac{i-1}{N}$ and $\frac{i}{N}$, include individual *i* in the sample. Another way to express this is as follows. Let *R* denote one of the random numbers generated. Include individual *i* in the sequence if

$$i-1 < R \times N \le i\,.$$

Since this could happen more than once for the same individual, this will be a sample with replacement. If you want a sample without replacement, ignore any repetitions and keep generating random numbers until you have *n* distinct members of the population.

Exercise: Suppose you have 20 students on your class roll and you want a random sample of 5 of them. Use your calculator and the procedure described above to do it.

With a spreadsheet program it is very easy to select a random sample of the rows in a worksheet. Insert a new column in the worksheet that contains a sequence of numbers generated by the program's random number generator. The numerical order of the entries in this column will be completely random. Sort the entire worksheet by the entries in this column. Now the new column will be in increasing numerical order but the original order of the rows will have been completely randomized. Select the first n rows. This is your random sample from the population represented by the rows of the worksheet.

Exercise: Open the teacher pay data set with a spreadsheet program. Calculate the population median and the *IQR* of the variable *Spend*. Use the procedure above to select a random sample of about 10% of the rows in the worksheet. Calculate the sample median for the variable for the variable *Spend*. Do this about 10 times and record all the sample medians. Find the median of all the sample medians and *IQR* of the sample medians. How do these compare to the population median and *IQR* of *Spend*?