Figure 8.10. PLANNED DATA COLLECTING: Experiments and Sample Surveys

Program 14 in: *Against All Odds: Inside Statistics*

Although simple random sampling (or equiprobable selecting) is the foundation of sampling designs (or sampling protocols), sample surveys in practice usually use more complex designs. This program begins with a look at these designs. In *stratified random sampling* (or equiprobable selecting from a stratified population), the population is divided into *strata*, groups of units or elements that are similar in some way important to the reponse. A separate SRS is then selected from each stratum. Stratified sampling designs are the analogue of block designs for an experiment. The video illustrates stratified sampling by a survey of recreational fishermen carried out by the National Marine Fisheries Service. The strata in this case are different locations for fishing.

National samples are *multistage samples* that select successively smaller regions within the population in stages. Each stage may employ SRS, stratified sampling, or another type of sampling. The General Social Survey conducted by the National Opinion Research Center is an example of a national sample that provides important information about public sentiment.

Failure to use probability sampling often results in *bias*, or systematic errors in the way the sample represents the population. This program also reminds us of some of the practical difficulties that may cause bias when sampling *people*. A sampling design that systematically misses part of the population, like the *Literary Digest* election poll mentioned in the video, will suffer from bias due to this *undercoverage*. Misleading results can also be due to *non-response* (some subjects can't be contacted or refuse to answer), to badly worded questions, or even to the race, sex, or appearance of the interviewer.

The deliberate use of chance in producing data is intended to eliminate bias. It also makes the outcomes of an experiment or sample survey subject to the laws of probability. The next four Programs will introduce probability; the final portion of this program takes a first look at the random behaviour of statistical data. First, some basic vocabulary: A number that describes a population is called a *parameter.* A number than can be computed from the data is called a *statistic.* The purpose of sampling or experimentation is usually to use statistics to make statements about unknown parameters.

A statistic from a probability sample or randomized experiment will not take the same value if the sampling or experiment is repeated. The *sampling distribution* describes how that statistic varies in repeated data collection. In the video you can watch a sampling distribution build up under repeated SRS from a population of beads. Formal statistical inference is based on the sampling distributions of statistics.

Bias, which we informally described as "favoritism", can be described more exactly in terms of the sampling distribution. Bias means that the centre of the sampling distribution is not equal to the true value of the parameter. We would like to use a statistic that has small bias and also does not vary greatly in repeated sampling or experimenting. The *variability* of the statistic is described by the spread of its sampling distribution. If the sampling distribution is normal, the standard deviation of the distribution describes the variability of the statistic. Properly chosen statistics from randomized data production designs have no bias due to selecting the sample or assigning the experimental units to treatments. The variability of the statistic is determined by the size of the sample or of the experimental groups: *larger* samples give *less* variable results. Notice in particular that as long as the population is much larger than the sample, the variability of sample statistics is influenced *only* by the size of the sample and *not* by the size of the population.

$\boxed{1}$ Using the terminology of our Course Materials, rewrite the following phrases or sentences from the video summary above.

- *A separate SRS is then selected from each stratum.* (First paragraph)
- *The deliberate use of chance in producing data is intended to eleminate bias.* (Fourth paragraph)
- *It also makes the outcomes of an experiment or sample survey subject to the laws of probability.* (Fourth paragraph)
- *..... a first look at the random behaviour of statistical data.* (Fourth paragraph)
- *A number that describes a population is called a* parameter. (Fourth paragraph)
- *A number that can be computed from the data is called a* statistic. (Fourth paragraph)
- *The purpose of sampling or experimentation is usually to use statsitcs to make statements about unknown parameters.* (Fourth paragraph)
- *A statistic from a probability sample or randomized experiment will not take the same value if the sampling or experiment is repeated.* (Fifth paragraph)
- *Properly chosen statistics from randomized data production designs have no bias due to selecting the sample or assigning the experimental units to treatments.* (Last paragraph) [this statement is shortened to ... *random sampling eliminates the bias* ... about 20 minutes into the video commentary.]
- *..... as long as the population is much larger than the sample, the variability of sample statistics is influenced* only *by the size of the sample and* not *by the size of the population.* (Last paragraph)

(*continued overleaf*)

$\boxed{2}$ Identify the three sources of *error* (in our terminology) mentioned in the third paragraph overleaf on page 8.47: ..... *Failure to use probability sampling* ....

- ● Rewrite the paragraph using our terminology.

1995-04-20