Figure 3.13. PLANNED DATA COLLECTING: Experiments and Sample Surveys

Program 13 in: *Against All Odds: Inside Statistics*

This program has two main purposes. First, we look at some more complex designs for experiments, especially the use of blocking. *Blocks* are groups of experimental units or subjects that are similar in some way important to the response. In a randomized block design, randomization is carried out separately within each block. Forming blocks is another type of con- trol in an experiment. The video shows an agricultural experiment in which the experimental units are small plots of land and the treatments are different *varieties* of strawberries. Because land in different locations varies in fertility, drainage and other ways that affect plant growth, it is helpful to divide the land into blocks containing a few plots in the same location. The differ- ent varieties are then assigned at random to plots separately within each block. In addition to blocking, the program also mentions experiments with more than one factor. In such experiments, a treatment combines a particular value of each factor.

The second purpose of this program is to introduce the design of sample surveys. In a sample survey, a *sample* is selected from the *population* of all people or things about which we desire information. Conclusions about the population are based on examination of the sample. The video contrasts a sample with a *census* which attempts to examine *every* member of the population. Even a census is usually not completely accurate, as a discussion of the U.S. population census shows. Some people are missed by the census, and this undercount is concentrated in the minority neighbourhoods of large cities. Only a census can give detailed information about every part of a population, but for overall information a sample is usually faster, less expensive, and just as accurate. The varied uses of sampling are shown by a look at the manufacture of potato chips: sampling is used at every stage, from the arrival of potatoes at the plant to checking the final product on supermarket shelves.

The *design* of a sample refers to the method used to select the sample from the population. The most important statistical principle in sampling design is to select the sample *at random* from the population. The basic probability sample is a *simple random sample*, which gives every possible sample of a given size the same chance to be the sample selected. Simple random samples are chosen by labelling the members of the population and using random digits to select the sample. Random sampling avoids the systematic favoritism or *bias* that often results when a sample is formed by human choice.

1 Comment critically on the last two sentences of the video summary given above.

- *Simple random samples are chosen by labelling the members of the population and using random digits to select the sample.*
- *Random sampling avoids the systematic favoritism or* **bias** *that often results when a sample is selected by human choice.*

When viewing this Program 13, keep in mind the following clarifications or details of the discussion.

Blocking in an *experimental* Plan means forming groups of units (the **blocks**) with the *same* values of one or more non-focal explanatory variates; units within a block are then assigned *different* values of the *focal* variate.

When answering a Question with a *causative* aspect, statistical best practice to manage comparison error (to reduce the limitation it imposes on the Answer) is to:

* block (to the extent that is feasible in the Question context) on known and measured lurking variates,
* use equiprobable assigning (EPA) to manage unblocked, unmeasured and unknown lurking variates,

Thus, blocking is used to manage the magnitude of comparison error arising from *known* sources of variation among population elements (‘experimental subjects’ or ‘test conditions’).

For commercial growers, important strawberry plant variates are:

* plant yield;	* berry colour;	* berry texture;	* berry toughness of skin;
* uniformity of berry appearance;	* berry firmness;	* berry aromatic properties;	* berry flavour.

Discussion of the U.S. population census in Program 13 reminds us of the importance of rigorously maintaining the distinction between the population and the sample.

* In the context on managing non-response error, it is of interest that 83% of census questionnaires are returned by mail.
* Discussion of the census *undercount* reminds us of the need in data-based investigating for unwavering attention to measurement error to manage the limitation it imposes on Answers.
* These Course Materials use **bias** as the *model* quantity representing (real-world) **inaccuracy**; both refer to behaviour under repetition, *not* the individual case.

About 18 minutes into the video, we should be careful to distinguish:

* *voluntary response*, where a respondent freely chooses to provide a response (the usual state of affairs in Canada, and which may raise matters involving *measurement* error), FROM:

(*continued overleaf*)

∗ *volunteer selecting*, which means having a sample made up of units who *choose* to participate; in the case of the Hite report, this is fewer than 5% of the roughly 100,000 women who received the questionnaire, resulting in a *severe* limitation imposed on Answers by *non-response* error.

About 23 minutes into the video, the '*three* samples of potatoes' is actually *one* sample of three potatoes.

∗ In the context of statistical process control (SPC) and *destructive* measuring, the five 30-pound samples of potatoes, comprising 150 pounds in total, were about 0.3% of the 45,000 pounds of potatoes in the truck load.

The appealing intuitive idea of a 'representative sample' – one that 'looks like' the (respondent) population with respect to the attribute(s) of interest – is equivocal statistically for four reasons:

∗ a sample selected by EPS is unlikely to be 'representative' in the sense just given for *all* attributes of potential interest – for instance, a sample may have small [possibly (close to) zero] sample error for estimating $\overline{\mathbf{Y}}$ but large sample error for estimating $\mathbf{S}$;
∗ the sample, of itself, provides no information about its 'representativeness';
∗ there is no selecting process known to yield a 'representative' sample, except taking a census;
∗ the terminology tends to obscure the distinction between the individual case (the *particular* sample) and behaviour under repetition (the properties of the selecting *process*).

Less equivocal terminology is *representative sampling*, with its implication of a selecting *process* (like EPS) which, in conjunction with adequate replicating, provides for quantifying sampling imprecision and so allows a particular investigation to obtain an Answer with acceptable limitation (in the Question context) due to sample error. However, these Course Materials *avoid* the terms 'representative' and 'representativeness' when referring to a sample (or a sampling protocol).

- Kruskal and Mosteller devote 50 pages to discussing the (sometimes ill-defined) meanings in statistical contexts of **representative sampling** in three articles in the *International Statistical Review*, **47**, 13-24, 111-127, 245-265 (1979). [UW Library call number HA 11.I505]