1: Measurement and Sampling

Introduction | Measurement | Population and Sample

Introduction

Statistics is *more than a compilation of computational techniques*; it is a means of *learning* from data, a *servant* of all science (Neymann, 1955), and a way of collecting, organizating, and interpreting numerical data starting with the planning of a study and ending with the presentation of final results.

The job of the statistician is as a combined *Data Detective* and *Data Judge*. The Data Detective uncovers patterns and clues, while the Data Judge adjudicates whether uncovered patterns are valid and can be generalized. Both detection and adjudication are important in the practice of statistics (Tukey, 1969). To concentrate on detection without adjudication would be an obvious mistake, for the facts need to be confirmed. On the other hand, to submerge detection to an inferior role would be erroneous, for where does new knowledge come from if not from detection?

Our reliance on statistics must be examined against an understanding of empiricism (as an idea, not as a doctrine), in the sense that attitude that beliefs are to be accepted and acted upon only if they first have been confirmed by actual experience. This broad definition accords with the derivation of the name from the Greek word *empeiria*, meaning "experience." The value of statistics as an aid to empiricism is that statistics seeks to make empiric processes more *objective* (so that things are observed as they are, without falsifying them to accord with some preconceived world view) and *reproducible* **(**so that we might judge things in terms of the degree to which observations might be repeated). The cornerstone of this process is measurement with overlapping stages of reasoning, summarized as follows:

- **Observation,** in which the investigator observes what is happening, collects information, and studies facts relevant to the problem. In this stage, statistics suggests what can most advantageously be observed and how data might be collected.
- **Hypothesis,** in which the investigator puts forth educated hunches or explanations for observed findings and facts. In this stage, the statistician helps format observations in a form that are comprehensible and understandable.
- **Prediction**, in which anticipatory deductions based on hypotheses are put forward in testable ways. Statistics can help only a little at this stage, for predictive insights are often intuitive and creative, such as was the case with "Newton's apple".
- **Verification**, in which data are collected to test predictions. In judging the extent to which predictions are borne out by observation, we recognize that data and predictions almost never agree exactly, even when theories are correct.

Measurement

Types of Measurements

Measurement is the assigning of numbers or codes according to prior-set rules.[*] It is how we get the numbers upon which we perform statistical operations. There are many ways to classify measurements. SPSS classifies measurements as either nominal, ordinal, or scale.

Nominal variables are named categories or attributes. For example, `SEX` (male or female) is a nominal variable, and so is `NAME` and `EYECOLOR`. Nominal variables are often referred to as **qualitative variables** or **categorical variables**.

Ordinal variables are rank-ordered characteristics and responses. For example, an `OPINION` graded 5 = strongly agree, 4 = agree, 3 = undecided, 2 = disagree, 1 = strongly disagree is ordinal. Notice that ordinal categories can be put in ascending or descending order, but difference ("distances") between possible responses are uneven. For example, the difference between 5 (strongly agree) and 4 (agree) is not the same as the difference between a 4 (agree) and 3 (undecided).

Scale variables represent quantitative measurements in which differences between possible responses are uniform. For example, `LENGTH` (measured in inches) is a scale measurement since the difference between 2 inches and 1 inch is the same as the difference between 3 inches and 2 inches. Scale variables are also called **quantitative variables** and **continuous variables**.

Notice that each step up the measurement scale hierarchy takes on the assumptions of the step below it and then adds another restriction. That is, nominal variables are named categories, ordinal variables are named categories that can be put in logical order, and scale variables are ordinal variables that have equal distances between ordered responses.

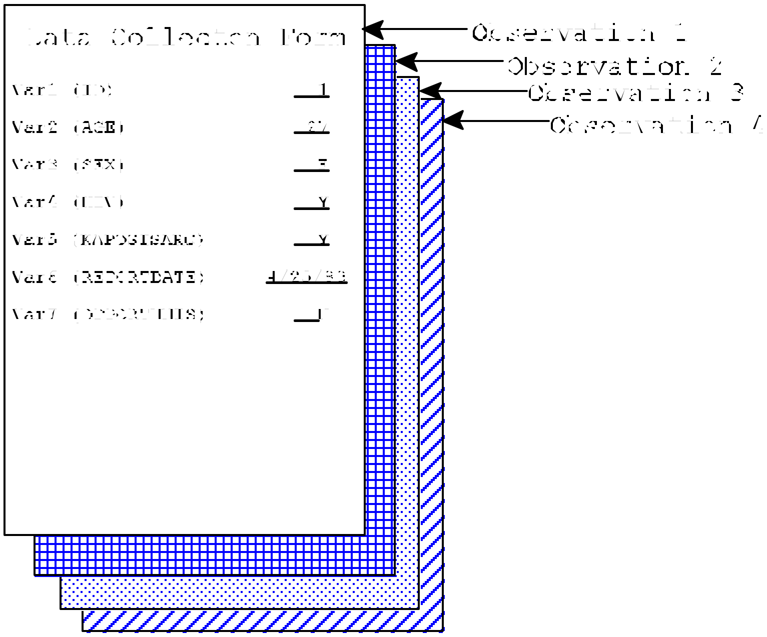
> ***Comment:*** Although distinctions between measurement scales appears to be straightforward, the lines between them occasionally get blurred. For example, an IQ score is usually considered to be a scale variable. However, strictly speaking, we have no assurance that the difference between an IQ of 70 and 80, for instance, means the same as the difference between an IQ of 80 and 90. Therefore, IQ scores can be thought of as either ordinal or scale. In fact, some statisticians believe that adhering to any type of measurement taxonomy can be misleading, and discourage the use of any single scheme.[†]

[*] Stevens, S. S. (1946). On the theory of scales of measurement. *Science 103*: 677 - 680.

[†] Velleman, P. F. & Wilkinson L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician*, 47, 65 - 72.

Data Structure

Regardless of whether data are experimental or nonexperimental, they are usually collected on a **data collection form**. Each completed data collection form represents information from a single **observation** (or, as SPSS calls it, a “case”), and each field or question on the data collection form corresponds to a **variable**:



Data Collection Form

Var1 (ID) 1

Var2 (AGE) 27

Var3 (SEX) F

Var4 (HIV) Y

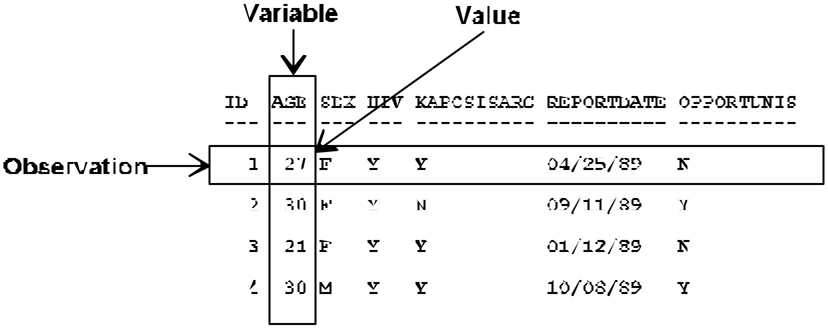
Var5 (KAPOSISARC) Y

Var6 (REPORTDATE) 4/25/89

Var7 (OPPORTUNIS) N

Notice that the **variables** represent measurements that can vary or be expressed as more than one value during the course of a study. For example, the above data collection form contains seven variables (ID, AGE, SEX, and so on). Thus, variables represent the generic “thing” being measured, and not any specific value or code for the measurement.

Once data are collected, they are organized to form a **data table**. Data tables are aligned so that each row contains data from a single observation and each column contains data from a single variable. The intersection of each row and column (each cell) contains a specific **value**:



ID	AGE	SEX	HIV	KAPOSISARC	REPORTDATE	OPPORTUNIS
1	27	F	Y	Y	04/25/89	N
2	30	F	Y	N	09/11/89	Y
3	21	F	Y	Y	01/12/89	N
4	30	M	Y	Y	10/08/89	Y

For example, in the above data table, the *value* of the AGE *variable* for *observation* 1 is “27.” The value of the variable OPPORTUNIS for observation 4 is “Y.”

Data Quality

"Garbage in and garbage out" (*GIGO*), or so the old saying goes. A statistical analysis is only as good as its data. Thus, we must place great emphasis on collecting good quality information. In addressing data quality we make the distinction between measurement error and processing error.

Measurement errors is differences between "true answers" and what appears on the data collection form. In surveys, measurement errors can be due to problems in questionnaire design and administration. Consider, for example, how subtle word choices can influence responses:

> Suppose I ask you to remember the word 'jam.' I can bias the way in which you encode and remember the word by preceding it with the word 'traffic' or 'strawberry.' If I have initially biased your interpretation of the word in the direction of traffic jam, you are much less likely to recognize the word subsequently if it is accompanied by the word 'raspberry,' which biases you toward the other meaning of jam. This effect occurs even though the subject knows full well that he is only supposed to remember the word 'jam' and not the contextual or biasing words. . . . We do not perceive or remember in a vacuum (Baddeley , 1999, p. 66).

Processing errors are errors that occur during data handling. Processing errors can occur at any level of data handling. Examples of processing errors include transpositions (e.g., 19 becomes 91 during data entry), copying errors (e.g., the number 0 becomes the letter O), data entry errors, and data programming errors. The most effective way to deal with processing errors is to identify the stage at which they occur and address the problem at that point. This may involve manual checks for completeness (e.g., checks for legible handwriting) or computerized checks during data entry (e.g., double entry and validation procedures).[*]

The study of data quality is a large topic, worhty of a course in itself. Still, here are some practical data quality "gotchas" for your consideration[†]:

- ✓ ALL data has errors (especially when the supplier insists it doesn't).
- ✓ Sometimes the person who supplies the data has already added two columns to get a third, not understanding that the computer can do a better job. If so, have the computer check her/his addition, and ask him/her to explain those that do not check.
- ✓ Do not use zero as a missing data code, and beware of this with data from other people.
- ✓ Once the data is in the machine, print it and verify the printed copy against the original data sheets.

[*] A separate chapter on data processing errors is presented later in *StatPrimer*.

[†] Source: Bill Knight, University of New Brunswick.

Population and Sample

Statistical studies are done to learn about the statistical characteristics of populations. This is accomplished by generalizing from a sample to a larger population of individuals. The sample is examined and the facts about it studied. Based on statistic calculated from data in the sample, inferences are made about the population.

Even when not sampling in a literal sense, data are conceptualized as a subset of a larger or super-population. This super-population (which is merely called "the population") is the collection of all possible values for a given variable. It may be real or hypothetical. Real populations have a finite number of possible values, while hypothetical value have an infinite or very large number of potential values, many of which may not be realizable. For now, let us consider sampling of a real population since this is easier to imagine.

Suppose we u want to learn about the AGE of a population. It may be possible to obtain information on the entire population. This is a **census**. However, conducting a census is often impractical and expensive, if not downright impossible. Therefore, the investigator usually chooses to study a subset of the population, which we refer to as the **sample**.

Over the past half-century, much has been learned about how to select a good sample. One thing that has been learned is that, whenever possible, a **probability sample** should be sought. A probability sample is a sample in which every population member has known probability of entering the sample, with samples drawn using chance mechanisms.[*]

The simplest type of probability sample is a simple random sample. A **simple random sample** is a sample in which each member of the population has an equal probability of entering the sample. For example, suppose we want to select a simple random sample from a population of 600 with individuals numbered 1 to 600. We let n represent the size of the sample and N represent the size of the population. The ratio n / N is the **sampling fraction**. For example, if we select a random sample of $n = 3$ in a population of $N = 600$, the sampling fraction = 3 / 600 = .005 (0.5%).[†] Thus, each person in the population should have a 0.5% chance of being selected . If some people in the population had more than a 0.5% chance of being selected (or if some had less), then the sample would *not* qualify as a simple random sample and would be prone to selection bias.

Sampling can be done with replacement or without replacement. **Sampling with replacement** is accomplished by "tossing" population members back into the mix after they have been selected. In this way, all N members of the population have an equal chance of being selected *at each draw*. In contrast, **sampling without replacement** is done so that once a population member has been drawn, this person is removed from further sampling. Thus, once a population member has been drawn, their subsequent probability of selection is zero and the probability that someone else is selected goes up a little. Introductory statistical procedures generally assume that the sample was done with replacement *or* that the population is so large in relation to the sample that this makes little difference.

[*] **Nonrandom samples** may be based on a systematic selection process or convenience. Such samples are prone to selection biases, thus presenting problems during analysis and interpretation.

[†] Lab 1 will demonstrate how to select a simple random sample.