

1: Measurement and Sampling

[Introduction](#) | [Measurement](#) | [Sampling](#)

Introduction

The theme of these notes is that statistics is *more* than just a compilation of computational techniques. Statistics is *not* merely pushing numbers through formulas or computers.

Statistics is about *learning* from data. It guides the way we collect, organize, and interpret numerical data. It helps us weigh evidence and draw conclusions.

The statistician is both a data detective and data judges (Tukey, 1969) The data detective uncovers patterns and clues, while the data judge decides whether clues can be trusted.

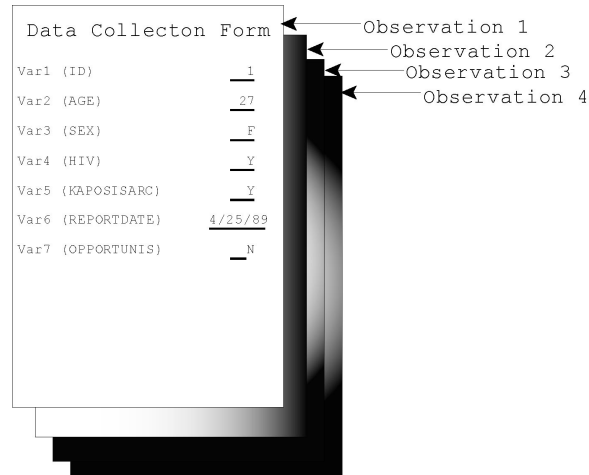
Statistics is the servant of all science (Neyman, 1955). The goal of statistics is to help science, and goal of science is to understand the nature of things. This begs the question as to what makes something scientific. While there are diverse views on this topic, Tukey (1962, p. 5) suggests these three constituents are essential to science:

- (1) Intellectual content
- (2) Organization of data into understandable forms
- (3) Reliance on tests of experience as the ultimate standard of validity

Statistics plays a constructive role in these three realms.

Data

Measurement is how we get data. It is “the assigning of numbers or codes according to prior-set rules” (Stevens, 1946). **Observations** are the individual units upon which measurements are made. You may think of an observation as the data on a single individual reported on a data collection form:



Data Collection Form	
Var1 (ID)	1
Var2 (AGE)	27
Var3 (SEX)	F
Var4 (HIV)	Y
Var5 (KAPOKISARC)	Y
Var6 (REPORTDATE)	4/25/89
Var7 (OPPORTUNIS)	N

Fig: DataCollectionForm.ai

The above figure shows four observations each with seven variables. **Variables** are the generic “thing” being measured. The above data collection form contains the following variables: ID, AGE, SEX, HIV, KAPOKISARC, REPORTDATE, AND OPPORTUNIS. (We use this FONT to represent variables.) A **value** is an individual measurement. Variables take on different values for different observations.

Data Table

Once collected, data are organized to form a **data table**. Typically, each row in a data table contains data from a single observation. Each column contains a single variable. Each cell contains a value. Here is an example of a data table.

	Variable						
	ID	AGE	SEX	HIV	KAPOSISARC	REPORTDATE	OPPORTUNIS
Observation	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

Fig: DataTable.ai

This data table has 7 variables (columns) and 4 observations (rows). There are a total of $7 \times 4 = 28$ values. The value of the `AGE` variable for observation 1 is 27. The value of the `OPPORTUNIS` variable for observation 4 is Y. (And so on.)

Measurement Scales

We consider three types of variables:

- **Categorical variables** represent named categories. Examples of categorical variables are `SEX` (male / female), `CASE` (case / non-case), and `EYE_COLOR` (brown, blue, other). Categorical variables are also called “nominal variables.”
- **Ordinal variables** represent rank-ordered categories. An example of an ordinal variable is `OPINION` graded 5 for “strongly agree,” 4 for “agree,” and so on. Another example is `STAGE` of cancer, grades 1, 2, 3, or 4. Ordinal data can be put in ascending or descending order, but differences between categories are *not* evenly spaced.
- **Quantitative variables** are numeric in the usual sense. The “distance” between quantitative values are evenly spaced. For example, an `AGE` of 2 has an equal distance from an age of 3 and an age of 1. Quantitative values can be placed on a number line. This allows us to perform arithmetic operations like summing and averaging on quantitative variables. Quantitative variables are also called “scale variables.”

Notice that each of the measurement scales — from categorical to ordinal to quantitative — takes on the assumptions of the step below and adds a further restriction. Ordinal variables are categories that can be ranked and quantitative variables are ordered measurements that have equal spacing between intervals.

Comment: Velleman and Wilkinson (1993)* point out that the distinction between categorical, ordinal, and quantitative measurements often gets blurred. For example, IQ scores are usually considered quantitative. However, strictly speaking, we have no assurance that the difference between an IQ of 70 and 80 means the same as the difference between an IQ of 80 and 90.

Data Quality

An analysis is only as good as the quality of the data upon which it is based. Fancy analyses cannot compensate for poor quality data. Statisticians have a saying for this: Garbage in, garbage out (*GIGO*).

The goal is to make observations **objective** (so that things are observed as they are, without shaping the data to conform to your own preconceived world view), **reproducible** (the degree to which the same value would be observed if measurement were taken repeatedly), and **valid** (measurement produces over the long haul will be correct).

In discussing data quality, we distinguish between measurement error and processing error.

Measurement error is the difference between “true answers” and what appears on the data collection form. **Processing errors** occur after data have been collected.

Measurement errors can be blatant or subtle. Consider how subtle word choices may influence responses to an interview:

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 can take many forms. Examples of processing errors include data transpositions (e.g., 19 becomes 91 during data entry), copying errors (e.g., the number 0 becomes the letter O), data entry errors, data programming errors, and so on. 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).

Finally, 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 to

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

† Source: Bill Knight, University of New Brunswick.

verify data input.

Sampling

The goal of statistics is to learn about *populations*. In statistics, the term population is used slightly differently than commonly conceived.

A statistical population is the set of all possible values for the variable.

Measurements for a given variable for each individual in a population would generate a *population of values* for that variable. For example, the ages of all people living in a particular region would represent the population for the variable AGE. We are not speaking of people, but of a large set of numbers.

Populations may be finite or infinite. We start by considering finite populations. Even finite populations are often too large to study in entirety. A subset of the population must be studied. This is a sample.

A sample is a subset of the population.

Over the past century, much has been learned about how to select a good sample. The goal is to make findings from the sample generalizable to the population. To achieve this goal, we should use chance mechanisms to select our sample. The most direct way to select a random sample is to select a simple random sample.

A simple random sample (SRS) is a sample in which each member of the population has an equal probability of entering the sample.

We will learn about the mechanics of selecting a SRS in lab. For now, let's just consider the general concept. The idea behind a SRS is that each person in the population has the same probability of entering the sample. Suppose, for example, you wanted to select a simple random sample of 6 individual from a population of 600. If this were a SRS, each person in the population would have to have a 6 in 600 (1%) chance of entering the sample. If anyone has a greater or lesser chance of entering the sample, the sample would *not* be a SRS. If we let n represent the sample size and N represent the population size. The ratio $\frac{n}{N}$ is the **sampling fraction**. Everyone in the population must have a n / N chance of entering the sample for it to be a SRS.

Sampling can be done with replacement or without replacement. **Sampling with replacement** is accomplished by “tossing” selected members back into the mix after they have been selected. In this way, any given individual can appear more than once in a sample. (All N members of the population have a n/N 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 selection.

Introductory statistical procedures generally assume the sample was done with replacement *or* the sampling fraction is so small that it makes little difference. Lab 1 will demonstrate simple random sampling.