1: Measurement and 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, and guiding the way we collect, organize, and interpret information.

The statistician is both a data detective and data judge (Tukey 1969). The detective collects evidence and uncovers clues. The judge decides whether clues can be trusted, weighs evidence, and helps us draw conclusions.

Data

Measurement is "the assigning of numbers or codes according to prior-set rules" (Stevens, 1946). It's how we get our data.

Observations are the units upon which measurements are made. Observation correspond to individuals (individual people, individual institutions, individual biological samples, etc.).

Data are collected on forms. Here's an illustration of **data collection forms** for four sequential observations:

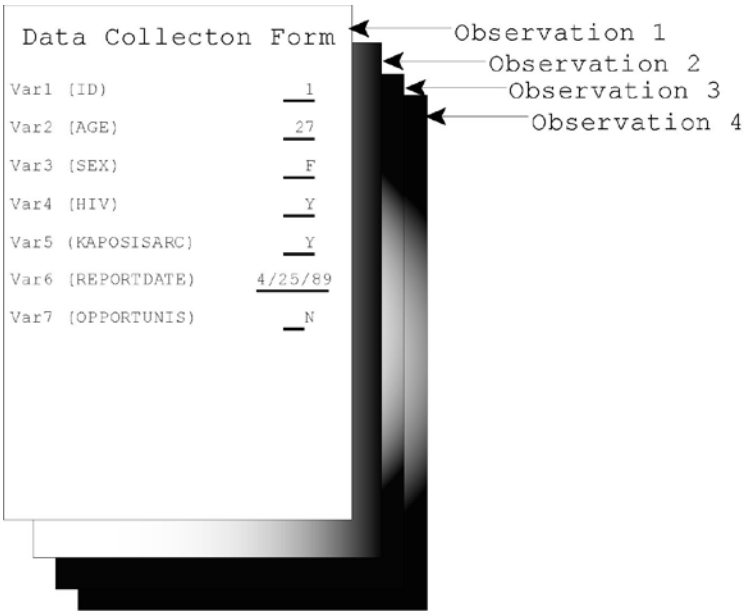
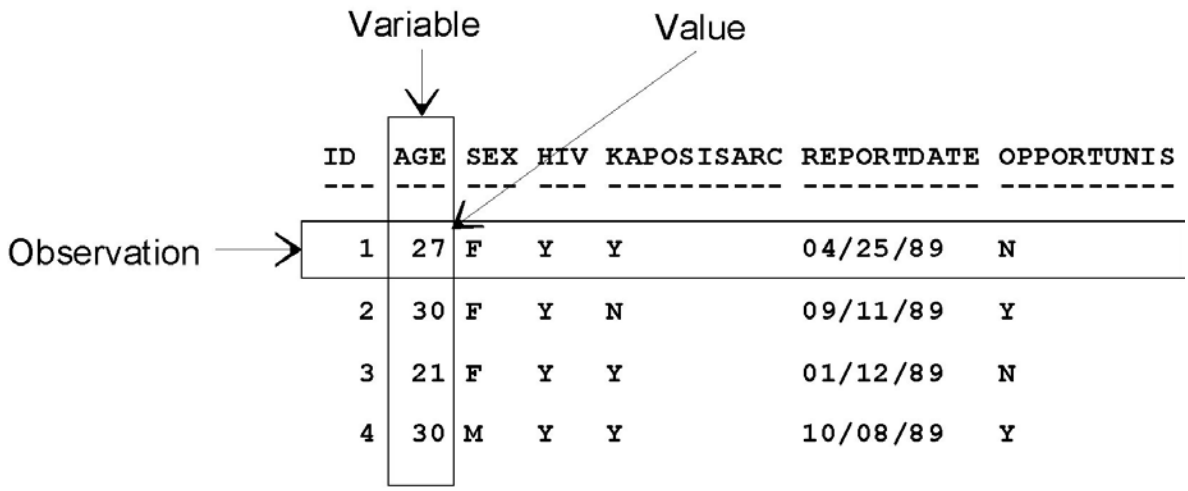


Fig: DataCollectionForm.ai

There are 7 **variables** on this form (ID, AGE, SEX, HIV, KAPOSISARC, REPORTDATE, and OPPORTUNIS). Do *not* confuse the term "variable" with "value." **Variables** can take on different values.
A **value** is a particular "outcome" for a variable. For example, the first observation above has a value of "F" (code for female) for the variable SEX.

Data Table

Data are compiled into **data tables**. Each **row** in a data table contains data from a single *observation*, each **column** contains data from a single *variable*, and each **cell** contains a single *value*. Here is an example of a data table.



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

Fig: DataTable.ai

This data table has 7 variables, 4 observations, and 28 values. The value of the variable `AGE` for observation #1 is 27. The value of the variable `OPPORTUNIS` for observation 4 is Y. (And so on.)

Measurement Scales

We consider three types of **measurement scales** for variables. These are:

- **Categorical variables** are named or "nominal" categories. Examples of categorical variables are `SEX` (male / female), `CASE` (case / non-case), and `EYE_COLOR` (brown, blue, other).

- **Ordinal variables** represent rank-ordered categories. An example of an ordinal variable is `OPINION` ranked from 5 for "strongly agree" 4 for "agree, 3 for "neutral" and so on. Another example is `STAGE` of cancer graded 1 for the least virulent, 2 for slightly more virulent, and so on. Ordinal data can be put in ascending or descending order, but differences between categories are *not* evenly spaced.

- **Quantitative variables** also called "continuous data" have true numeric meaning. This means the "distance" between values measures a quantity. For example, an `AGE` of 2 is one more than an age of 3 and is twice as much as an age of 1. Quantitative values can be logically placed on a number line. This permits for arithmetic operations like summing and averaging.

Note that measurement scales at each step, from categorical to ordinal to quantitative, take on a further restriction. Ordinal variables are categorical that can be ranked. Quantitative variables are ordered that have quantitatively spacing between intervals.

Ordinal data can be treated as categorically or quantitatively, depending on further assumptions the researcher is willing to venture.

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 objective of statistics 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), **accurate** (the extent to which the results of the measurement agrees with the true nature of things), and **precise** (the extent to which a measurement is repeatable or reproducible).

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 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).

Sampling

The goal of statistics is to learn about *populations*.

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

The statistical term **population** is different than the demographic term. We are not talking about a group of people (necessarily). We are talking about a *population of values*.

For example, if we are discussing HDL cholesterol levels in men who take testosterone therapy, we are not talking about the men. We are talking their HDL cholesterol values. And we are normally talking about all men who might supplement with testosterone (a potential set of values).

Statistical populations are often too large or too hypothetical to study. We must there **sample** the population.

A **sample** is a subset of the population.

To make findings from a sample generalizable, the sample must be selected in specific ways. The most direct way to achieve a scientifically generalizable sample is via a simple random sample (SRS). A **simple random sample (SRS)** is a sample in which each member of the population has an equal chance to enter the sample: every conceivable sample of size *n* taken from the population has the same probability of selection.

In an SRS of 6 individual from a population of 600, each potential value has a 6 in 600 (1%) chance of entering the sample. If anyone has any different chance of entering the sample, the sample would *not* be a SRS. If a sample is systematically excluded from the sample, the sample is not random.

Let *n* represent the sample size. Let *N* represent the population size. The ratio of *n* to *N* is the **sampling fraction**. In an SRS, everyone should have a *n* / *N* chance of entering the sample.

Sampling can get complicated. For example, in a finite population, 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. Sampling can first divide (stratify) the population into subgroups before random selection (stratified sample), can sample cluster (cluster sample), and can occur in multiple stage (multi-stage sampling. These more advance sampling techniques are beyond the scope of introductory statistics. Almost all the techniques introduced in introductory statistical classes are based on simple random samples or some variant of a simple random sample.