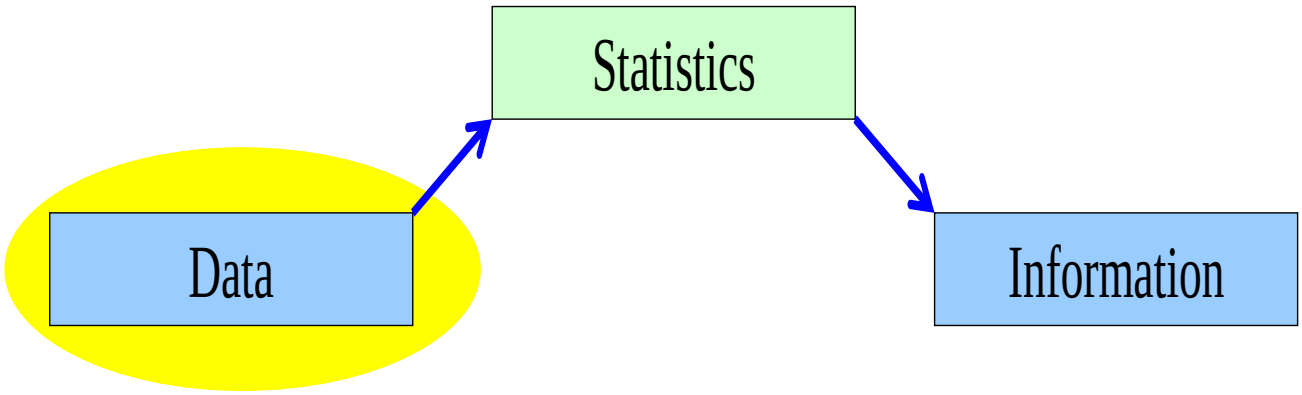
Data Collection and Sampling

OPRE 6301

Recall. . .

Statistics is a tool for converting *data* into *information*:



But where then does *data* come from? How is it gathered? How do we ensure its accurate? Is the data reliable? Is it representative of the population from which it was drawn? We now explore some of these issues.

Methods of Collecting Data...

There are many methods used to collect or obtain data for statistical analysis. Three of the most popular methods are:

- Direct Observation
- Experiments, and
- Surveys.

Surveys...

A *survey* solicits information from people; e.g. Gallup polls; pre-election polls; marketing surveys.

The *Response Rate* (i.e. the proportion of all people selected who complete the survey) is a key survey parameter.

Surveys may be administered in a variety of ways, e.g.

- Personal Interview,
- Telephone Interview, and
- Self-Administered Questionnaire.

Questionnaire Design

Over the years, a lot of thought has been put into the science of the design of survey questions. Key design principles:

1. Keep the questionnaire as short as possible.
2. Ask short, simple, and clearly worded questions.
3. Start with demographic questions to help respondents get started comfortably.
4. Use dichotomous (yes | no) and multiple choice questions.
5. Use open-ended questions cautiously.
6. Avoid using leading-questions.
7. Pretest a questionnaire on a small number of people.
8. Think about the way you intend to use the collected data when preparing the questionnaire.

Sampling. . .

Recall that statistical inference permits us to draw conclusions about a population based on a sample.

Sampling (i.e. selecting a sub-set of a whole population) is often done for reasons of *cost* (it's less expensive to sample 1,000 television viewers than 100 million TV viewers) and *practicality* (e.g. performing a crash test on every automobile produced is impractical).

In any case, the *sampled population* and the *target population* should be **similar** to one another.

Sampling Plans. . .

A *sampling plan* is just a method or procedure for specifying how a sample will be taken from a population.

We will focus our attention on these three methods:

- Simple Random Sampling,
- Stratified Random Sampling, and
- Cluster Sampling.

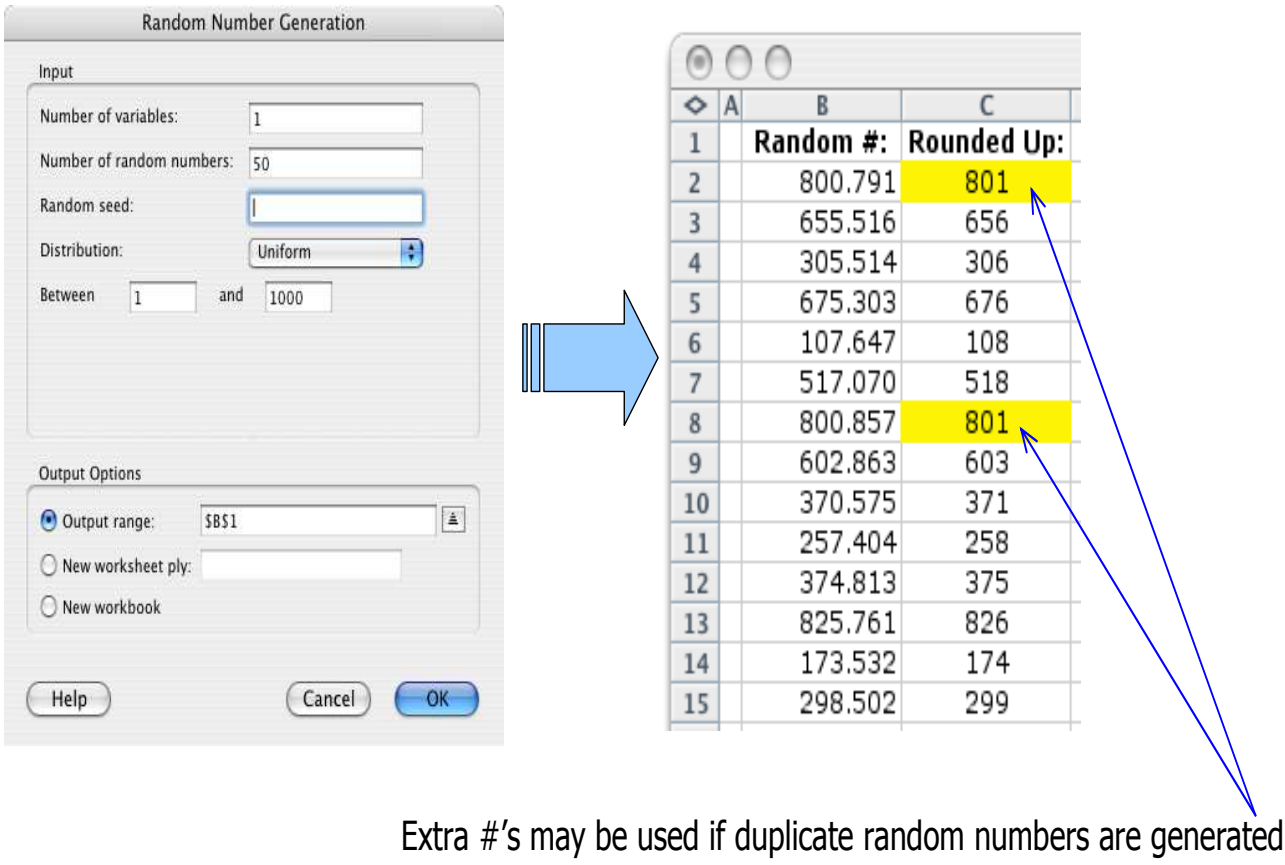
Details. . .

Simple Random Sampling...

A *simple random sample* is a sample selected in such a way that every possible sample of the same size is equally likely to be chosen.

Drawing three names from a hat containing all the names of the students in the class is an example of a simple random sample: any group of three names is as equally likely as picking any other group of three names.

Example: A government income tax auditor must choose a sample of 40 (usually denoted by n) of 1,000 (usually denoted by N) returns to audit...



The Excel file "C5-01-Random_Sampling.xls" demonstrates how to use the Excel function RAND() to generate a random sample from a population. Detailed explanations are provided in the spreadsheet itself.

Stratified Random Sampling. . .

A *stratified random sample* is obtained by separating the population into *mutually exclusive* sets, or strata, and then drawing simple random samples from each stratum.

Strata 1 : Gender	Strata 2 : Age	Strata 3 : Occupation
Male	< 20	professional
Female	20-30	clerical
	31-40	blue collar
	41-50	other
	51-60	
	> 60	

We can acquire about the total population,
make inferences within a stratum
or make comparisons across strata

After the population has been stratified, we can use *simple random sampling* to generate the complete sample:

Income Category	Population Proportion	Sample Size n = 400	n = 1000
under $25,000	25%	100	250
$25,000 - $39,999	40%	160	400
$40,000 – $60,000	30%	120	300
over $60,000	5%	20	50

If we only have sufficient resources to sample 400 people total, we would draw 100 of them from the low income group…

…if we are sampling 1000 people, we'd draw 50 of them from the high income group.

Cluster Sampling...

A *cluster sample* is a simple random sample of groups or clusters of elements (vs. a simple random sample of individual objects).

This method is useful when it is difficult or costly to develop a complete list of the population members or when the population elements are widely dispersed geographically.

Cluster sampling may increase sampling error due to similarities among cluster members.

Sample Size...

This is an important issue. Numerical techniques for determining sample sizes will be described later, but suffice it to say that the larger the sample size is, the more accurate we can expect the sample estimates to be.

Sampling and Non-Sampling Errors. . .

Two major types of error can arise when a sample of observations is taken from a population: *sampling error* and *non-sampling error*.

Sampling error refers to differences between the sample and the population that exist only because of the observations that happened to be selected for the sample.

Non-sampling errors are more serious and are due to mistakes made in the acquisition of data or due to the sample observations being selected improperly.

Details. . .

Sampling Error...

Sampling error refers to differences between the sample and the population that exist only because of the observations that happened to be selected for the sample.

Another way to look at this is: the differences in results for different samples (of the same size) is due to sampling error:

E.g. Two samples of size 10 of 1,000 households. If we happened to get the highest income level data points in our first sample and all the lowest income levels in the second, this is a consequence of sampling error.

Increasing the sample size *will* reduce this type of error.

Non-Sampling Error...

Non-sampling error are more serious and are due to mistakes made in the acquisition of data or due to the sample observations being selected improperly.

There are three types of non-sampling errors:

- Errors in data acquisition,
- Nonresponse errors, and
- Selection bias.

Increasing the sample size will *not* reduce this type of error.

Details...

Errors in Data Acquisition

. . . arises from the recording of incorrect responses, due to:

— incorrect measurements being taken because of faulty equipment,

— mistakes made during transcription from primary sources,

— inaccurate recording of data due to misinterpretation of terms, or

— inaccurate responses to questions concerning sensitive issues.

Nonresponse Error

. . . refers to error (or **bias**) introduced when responses are not obtained from some members of the sample, i.e. the sample observations that are collected may not be representative of the target population.

As mentioned earlier, the *Response Rate* (i.e. the proportion of all people selected who complete the survey) is a key survey parameter and helps in the understanding in the validity of the survey and sources of nonresponse error.

Selection Bias

. . . occurs when the sampling plan is such that some members of the target population cannot possibly be selected for inclusion in the sample.