Sampling Methods and Sampling Size

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? This chapter explores 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,
- Self Administered Questionnaire, and
- Internet

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.

Types of sampling

- Non-probability samples
- Probability samples

Non probability samples

- Convenience samples (ease of access)

  sample is selected from elements of a population that are easily accessible

- Snowball sampling (friend of friend….etc.)

- Purposive sampling (judgemental)

  You chose who you think should be in the study

Cheaper- but unable to generalise

potential for bias

Probability samples

- Random sampling

  Each subject has a known probability of being selected

- Allows application of statistical sampling theory to results to:
  - Generalise
  - Test hypotheses

- Probability samples are the best

  Ensure Representativeness and Precision

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.

- Random sampling,by far, is the most common one used.

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.

VERY EASY TO DEFINE!

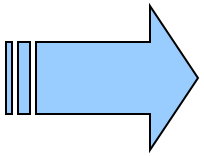
VERY, VERY DIFFICULT TO DO!

- Random sample of 100 cokes bottles today at the coke plant.
- Random sample of 50 pine trees in a 1000 acre forest.
- Random sample of 5 deer in a national forest.

Simple Random Sampling…

A government income tax auditor must choose a sample of 5 of 11 returns to audit…[Can do many different ways]

Person	Generate Random #
baker	0.87487
george	0.89068
ralph	0.11597
mary	0.58635
sally	0.34346
joe	0.24662
andrea	0.47609
mark	0.08350
greg	0.53542
aaron	0.37239
kim	0.73809



	Person	Sorted Random #
1	mark	0.08350
2	ralph	0.11597
3	joe	0.24662
4	sally	0.34346
5	aaron	0.37239
	andrea	0.47609
	greg	0.53542
	mary	0.58635
	kim	0.73809
	baker	0.87487
	george	0.89068

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

Stratified Random Sampling…

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	Sample Size 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…

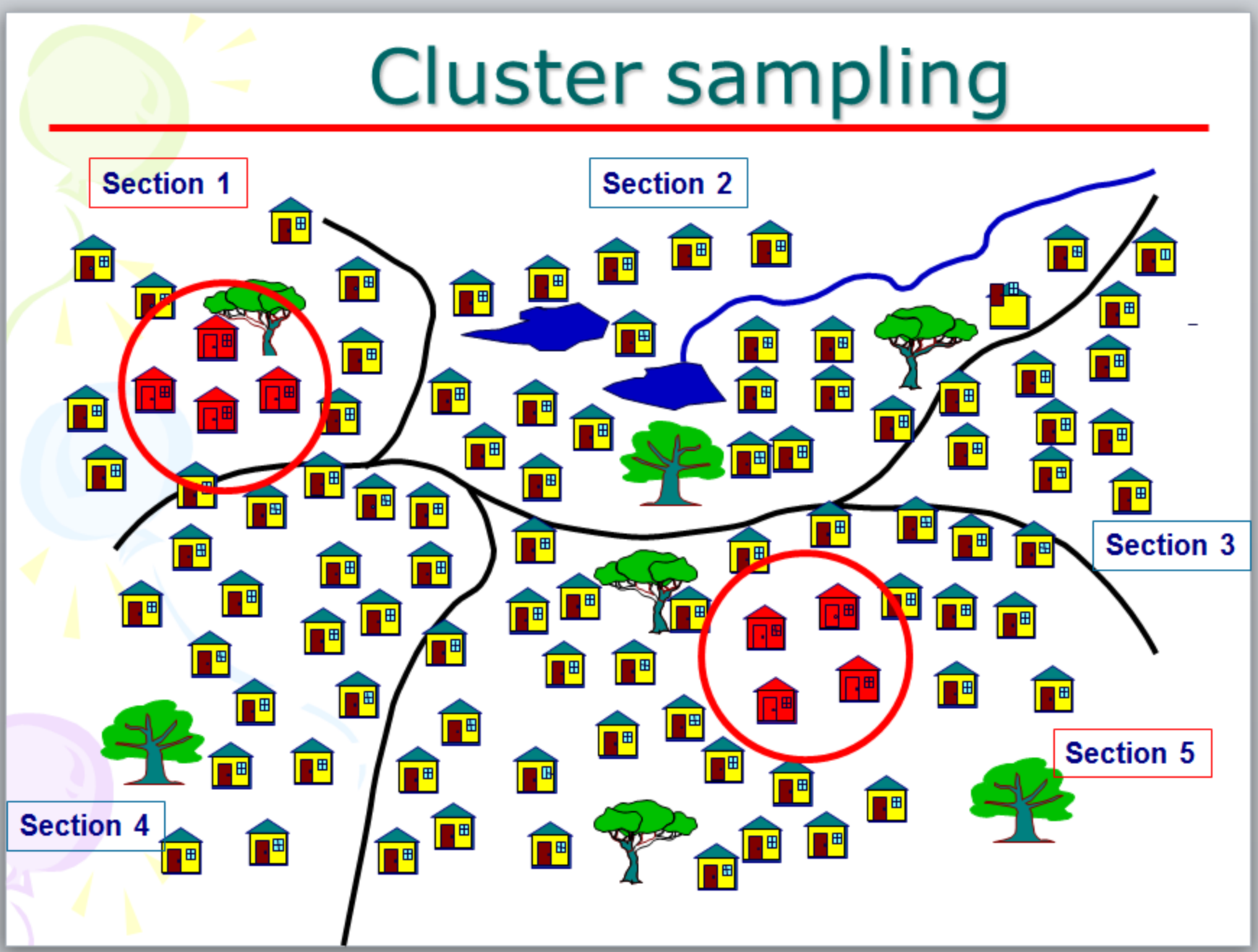
Cluster: a group of sampling units close to each other i.e. crowding together in the same area or neighborhood

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. Used more in the "old days".

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

Cluster sampling



Sample Size...

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 ***nonsampling 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. Random and we have no control over.

Nonsampling errors are more serious and are due to mistakes made in the acquisition of data or due to the sample observations being selected improperly. Most likely caused be poor planning, sloppy work

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.

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

Nonsampling Error...

Nonsampling errors are more serious and are due to mistakes made in the acquisition of data or due to the sample observations being selected improperly. Three types of nonsampling errors:

Errors in data acquisition,

Nonresponse errors, and

Selection bias.

Note: increasing the sample size **will not** reduce this type of error.

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.

Sampling Size

For descriptive statistics, to have 95% confidence level in estimating population parameters using a sample, can use:

1. Krejcie and Morgan (1970) Table. (Pg 295, Sekaran and Bougie).

2. Bartlett's Table

Bartlett, J.E., Kotrlik, J.W., Higgins, C.C. (2001). Determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal*, 19 (1), pp. 43-50.

http://api.ning.com/files/dDaMclZ3KnGUTT6nb2fPTHLjun-LLZEfrgdcoswcvTsB60CmiaZ93cmYBLFd1wUyFBUK4H9eT767qY8mUR7PWj88cc1Xw6h/SampleSizeDetermination.pdf

Sample Size for Inferential Statistical Analysis

Statistical power is the probability of not missing an effect, due to sampling error, when there really is an effect to be found.

Power is the probability (prob = 1 - β) of correctly rejecting H_o when it really is false.

Conventions And Decisions About Statistical Power

Acceptable risk of a Type II error is often set at 1 in 5, i.e., a probability of 0.2.

The conventionally uncontroversial value for "adequate" statistical power is therefore set at 1 - 0.2 = 0.8.

People often regard the minimum acceptable statistical power for a proposed study as being an 80% chance of an effect that really exists showing up as a significant finding.

Sample Size for Inferential Statistical Analysis can be determined using a software, GPower.

http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/download-and-register