**Sampling**

**Today’s**

* **Sampling**

**Intro**

*Question: What is sampling?*

*Question: Why is sampling necessary?*

A statistician wants to generalise about a class of individuals. This class is called *population.*

*Question: Give an example as why you would want to generalise about a population?*

In this session, our population of interest is votes casted by the population.

**Terminologies**

Parameter: Numerical facts about the population

Statistics: Numerical facts about the sample

Statistics are what we know; parameters are what we would like to know.

In forecasting a US presidential election the relevant parameters are

* Who are the eligible voters
* Percentage of eligible voters who have registered to vote
* What is their voting preference

**Intro to election polling**

*Question: What are election polls?*

*Question: What are the types of election polls?*

*Question: Which are the agencies that conduct election polls in India?*

*Question: How are they conducted? Say pre-poll and post poll predictions?*

*Question: Do you think this question – how an effective polling must be conducted – is a Statistical question? If so, why?*

… Because it involves sampling…

**Good Sampling**

*Question: What are the characteristics of a good sampling method?*

A sampling procedure should be fair. That means, selecting people for inclusion in the sample in an impartial way, so as to get a representative cross-section of the public.

*Question: What is selection bias?*

It is a systematic tendency on the part of the sampling procedure to exclude one kind of person or another from the sample.

*Question: What is non- response bias?*

People choosing not to respond to questionnaires.

**History**

In the US, initial polling used to be based on questionaries’ sent to people by mail whose name appeared on telephone books and club membership lists.

The polling results were often wrong.

*Question: Why do you think so?*

* It depends if the rich and poor voted differently
* If they didn’t, then this may not have been necessarily wrong

*Question: But there was 10 million Q sent? Do you think a large sample would eliminate sampling errors?*

Answer: No. Once biases are introduced during sample selection, the final results will not be accurate.

**Good Sampling Features**

* Every observation has an equal chance of being picked up.
* There is no bias.
* There is an element of probability in the process of selection.

**Setting Up a Nation-wide Poll**

*Question: How do you set up a nation-wide polling?*

* Divide country to different regions (Strata)

This is usually done as West, North, South and East.

*Question: What would you do next?*

* Make a list of locations with a high cut-off population

*Question: What do you do next? Can you go to every location selected above?*

* Use random selection and select only those locations that you have resources to station
* How do you make a random selection?

*Question: Within each town/city/village, where do you go to next?*

* Make a list of wards
* Make a random selection of wards

*Question: What is a ward? It consists of Polling Booths?*

* Make a list of booths
* Make a random selection of booths inside a ward

*Question: What next?*

* Make a list of households
* Randomly select households

*Question: Once you have households, what is the obvious next move?*

* Interview 18+ year olds in the household

**Types of Sampling**

*Stratified Sampling*

* Divide population into strata
* From each strata, randomly select items based on its proportion in the population

Strata are groups with the constituents being similar. While simple random sampling at the first stage is good, dividing a population into strata and selecting from each strata increases the information content.

Examples

Population – Male and Female

Household Income

Under $ 25000  
25000 – 50000  
Over 50000

We avoid stratifying when there is no connection between the survey and the strata. For example, little purpose is served in trying to determine whether people within religious strata have divergent opinions about the tax increase

*Cluster Sampling*

In most cases, it is difficult to stratify the population. We need to look for natural strata. Such as in the case of election polling, Wards and Polling Booths are the natural strata.

A sample of clusters could then be randomly selected, and every household within these clusters could be questioned to determine income

But true variation in the population may be lost if you rely heavily on cluster sampling.

*Systematic Sampling*

Assumes the population data is available in a list. A random starting point is selected and a random interval is selected. Elements starting from the starting point plus the interval are chosen. Example, testing every 5th car coming out of an assembly line.