Math 324 Lecture 2

Where are we at?

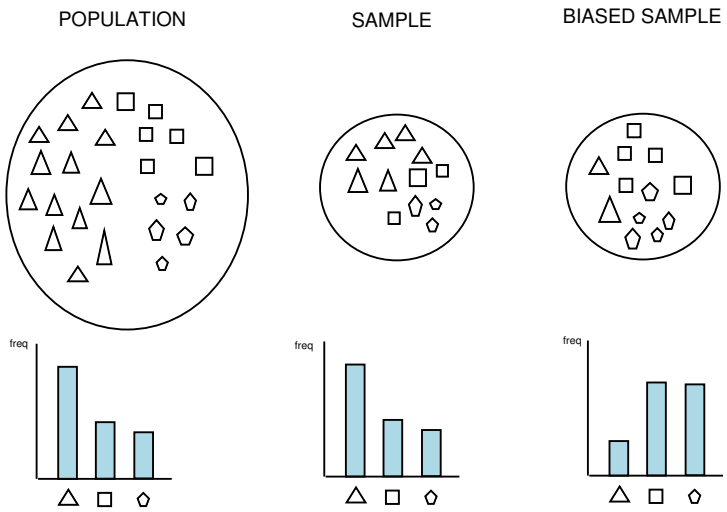
Last time we talked about data and how statistics was about collecting, organizing and understanding data. Today we will expand on these issues.

Populations and samples

First some definitions

- **Population** a well defined set of objects/units/individuals that we are interested in studying eg All adults in the USA, all widgets produced at a factory in a month, all bottles filled in a day at a soft drink company
- **Sample** a subset of the the population eg the first 1000 people selected by randomly choosing from valid social security numbers, every 10th widget off the assembly line, the bottles filled in the first hour of the day

A study of all the elements of a population is called a *census*. In an ideal world for anything which we wanted to study we could carryout a census, unfortunately this is almost never possible. Why? Cost or time prohibitive or perhaps our measurement method is destructive. So instead we usually instead look at data from samples from the population. One property that we would like is that our sample reflect the population from which it is drawn so that any conclusions we make based on the sample can be generalized to the population. We would call a sample that does not reflect the general population as biased.



How to sample from a population

How do you sample from a population without introducing bias? Well, this depends on a lot of things. First of all, what are we actually sampling from? The answer is what statisticians call the *sampling frame*. You can think of this as a list of all the possible objects or individuals from which the sample could be drawn.

The simplest method of selecting a sample from the sampling frame is to take a *simple random sample*. eg label every object with a distinct number from 1 to N. Then create tickets consecutively numbered from 1 to N and place in a big box, remember this is a thought experiment so we need not worry about things like how long it would take us to do this if N is really large. Shake the box for a long time so that all the tickets get really mixed up and then one by one draw out tickets from the box until with have a sample of desired size. Note that every individual has an equal chance of being selected into the sample. In practice we would probably use a computer to generate (pseudo-)random numbers.

Another common method of sampling is known as *stratified sampling*. Basically we divide the sampling frame into non overlapping groups and then take simple random samples from each one. In this manner we can ensure that certain subsets of the population are not over or under-represented in our sample.

Many times people will use a *convenience sample*. What does this mean?

Basically that we take a sample that is easy for us to get without systematic randomization. The problem with this is that this will often lead to a biased sample. For instance suppose we are interested in the average length (or weight) of fish in a lake. To sample we use a particular type of net that allows us to collect fish very quickly, but has a very coarse weave. In this case our sample of fish based upon what is in our net would be biased towards larger fish, since the smaller fish would slip right through the net.

Other things we should worry about in sampling include *under-coverage*, which is where our sampling frame does not contain all groups in the population, and *non-response* which is when an individual can not be contacted or does not wish to co-operate.

Some examples

- At a party there are 30 students 21 years old or older and 20 students under 21 years old. A research chooses 3 students from the 21 or older age group and 2 students from the under 21 age group to interview about their attitudes to alcohol. Does every student have equal probability of being selected? Is this a Simple Random Sample?

  Answer: The probability of selecting any one of the over 21 year olds is $\frac{3}{30} = \frac{1}{10}$ and the probability of selecting any one of the under 21 year olds is $\frac{2}{20} = \frac{1}{10}$ so every individual has equal probability of being selected. One of the conditions of a Simple Random Sample is that every individual has equal probability of being selected. However a Simple Random Sample would allow us to pick any 5 students (ie we could, by chance, pick 5 from the over 21 age group). Instead we refer to our sampling situation as Stratified Random Sampling. The two strata in this case are the age groups. Within each age group we have simple random samples.

- Telephone surveys. Which of the following two method should be preferred and why? Is either perfect.

  a A sample of households in a community is selected at random from the telephone directory.

  b After selecting an exchanges (or set of exchanges) the final four digits are selected at random.

Answer: The problem with (a) is that it misses people who have unlisted numbers. This introduces a bias, in particular we would call this an under-coverage bias. Our sample is missing part of the population (the set of all households in the community. (b) is better because we no longer miss people with unlisted numbers. However we also have a second under-coverage bias, in particular household with no telephones. Neither (a) or (b) would include these households.