Chapter 8 : Producing Data and Sampling

This chapter introduces the more ideas about how data is produced and what we wish to do with it.

Data can be observational or experimental. In an experiment, the experimenter (the one performing the experiment) imposes some condition or treatment and observes the response. In an observational study conditions are not imposed, either because it is not possible or ethics prevent this. Experimenters can study questions such as does some treatment prevent some disease; eg does aspirin prevent heart attacks. Experiments provide useful data, but only if properly designed.

- Observational study : responses or measurements are observed, but no attempt is made to influence the result
- experimental study : deliberate conditions are imposed and the response is observed. One tries to conclude if the condition (or changes in the condition) cause the response

Read example 8.1 (4th edition) discussing Hormone replacement therapy. By 1992 many national medical organizations, including Canada and the United States advised women to take hormone therapy after menopause. This was based on years of observational studies where it was observed that women taking hormone therapy had fewer health issues, including fewer heart attacks; the test mentions this reduction in risk is by 35% to 50%.

The women who took hormone replacement tended to be richer, better educated and was their doctors more often. Are they representative of all women? They tend to be healthier than the overall population of women.

From the period 1992 to 2002 several controlled trials were conducted. These were *randomized trials* in which a coin toss (or an equivalent randomization device) was used to assign women to either the hormone or *placebo* group. This randomization has the effect that it removes *selection bias* is how women enter the treatment group, in this case the hormone treatment group. In particular for women in the study education or wealth or overall health did not increase their chances of being in the treatment group. By 2002 the overwhelming evidence is that hormone treatment does not affect the risk of heart attack. It was an expensive treatment for no gain. It is now no longer used a suggested risk reduction treatment.

In general anecdotal evidence or observational studies do not help in understanding causal relations or effects. This is why promising drugs are required to undergo clinical randomized trials before allowing these to be used.

Sometimes these observational studies, when interpreted as causal can

have disastrous effects on populations.

Thalidomide was developed in the 1950s to prevent nausea during pregnancy. It was observed to be helpful in this. However it was not supported by clinical trials before common use. One side effect was that it was found to produce fetal limb abnormalities, in fact horrible types of abnormalities. A google search for "Image results for thalidomide" will turn up some pictures of such children, who numbered in the many thousands. A form of thalidomide is currently being studied as a treatment for aids and leprosy.

Confounding

Another problem with observational studies is *confounding*. We have seen this type of problem earlier in terms of regression and correlation. While we may observe some association or very good linear regression fit we should be careful of interpreting this as causation. There may be some common underlying phenomenon driving both sets of observations.

Example 8.2 (4th edition) gives one such example dealing with moderate drinking levels and health. We also saw an earlier example about polio and ice cream sales, in which there was an underlying temperature or summer condition that lead to both increased ice sales and spread of the polio virus. Another polio story is that before vaccines came into effect, children of relatively well off or wealthy families suffered more severe crippling and rates of polio than did children of less well off families. This was the opposite of what is observed in many other diseases, for example in TB (tuberculosis). Can you (the student) suggest why this might be?

Sampling

- Population : the population in a statistical study is the entire group of individuals or objects about which we want information
- Sample : a sample is a part (or subset) of the population from which we collect information or data
- Sampling design : this describes the method by which we choose a sample from the population

A Canadian census (complete enumeration of everyone in Canada) occurs every 5 years. Since it is expensive this is not done every year, but surveys or sampling of the population is done monthly. Also monthly employment surveys, various business and transportation surveys are carried out, many by Statistics Canada, and many by private sampling, survey or polling companies.

Bad sampling or survey designs are easy to obtain. Two common methods of sampling are convenience sampling and voluntary response samples. These can easily lead to *bias*, that is it systematically favours certain members of the population being chosen, and possibly excludes certain segments of the population. Effectively this changes the population from that desired to a different and possibly not even the correct one.

Example : Go to a shopping mall in mid afternoon to estimate the number of people employed. Since many jobs do not allow employed people to go to a shopping mall in a mid afternoon this part of the *statistical population* are excluded from being in the sample.

Example : give a medication only to healthy people and claim *people who take this medication are healthy*

A related idea is to study a coin. Flip the coin 10 times, but keep only those outcomes in which at least 5 come up heads. Report only these outcomes.

In these bad designs, one can have very misleading information at the conclusion of the sample and study. For example in the last example the average number of heads in 10 fair coin tosses is then about 6, much larger than the 5 that is expected for 10 fair coin tosses.

Simple Random Sample

A simple random sample is one in which every set of n members has the same chance of being selected.

For example if a population consists of 10 individuals then in a sample of size 3, every one of the possible sets of size 3 (there are 120 of these) has the same chance of being the sample.

How can one carry out a random sample? If the population consists of N individuals or members, then these should be numbered $1, 2, 3, \ldots, N$. Chose one of these at random. From amongst the remaining $N - 1$ chose a second one at random. Then from amongst the remaining $N - 2$ chose a third at random, etc until n are chosen.

How does one choose one item out of M at random? (M refers to N at the first stage, then $N-1$ at the second stage, then $N-2$ at the third stage, etcetera). One may put N identical slips of paper into a hat or drum (bingo drum), mix these up well and pull out one slip of paper. This works well when N is not too large, for example 5 or 10, but is not so easy to carry out if N is in the hundreds or thousands. There are also computer programs that can *simulate* this system so one does not have to cut or construct N identical slips of paper, write out N numbers, find a hat or drum, mix these well and then pull one out. There are also tables of Random Digits (eg Table B in the Moore text), but computer programs are now much more useful.

Other common sampling designs

- Probability Sampling : a sample is chosen by chance. This is useful when one wishes to sample with probability proportional to size, for example to sample cities or companies proportional to size. Large corporations will have a much larger chance of being included in a sample than a small company. In a simple random sample of corporations all companies have an equal chance of being chosen.

- Stratified Random Sample : In this type of design, the population is divided in strata or groups which are common in some important feature to us; for example large, medium or small corporations. Within each stratum a Simple Random Sample is chosen. These are useful as they ensure your sample will contain some of individuals from each of the strata.

  This type of sample is used by Statistics Canada in their monthly employment data collection. There provinces or regions are strata, as well small town and large cities in some provinces. Statistics Canada also has quarterly surveys on businesses, and has strata constructed so that various regions and industry classes are included in the sample.

  Stratified sampling is used by accountants to audit books of accounts payable or receivable. A complete enumeration of all accounts receivable would simple cost too much (time and money) to look at all small receivables. Typically all large receivables are included in an audit.

When a sample and its results are reported, the report should include

some information on how the sample was obtained. Without this information one cannot reasonably decide if the survey or sample results are *biased*, that is some outcomes are favoured and some are less likely to occur.

Things that can go wrong or make sampling interpretations difficult

Some main concerns are

- undercoverage : some groups in a population are not in a given sample. For example if one took a simple random sample of size 2 of the 10 provinces in Canada there is an 80% chance that Ontario is not included in the sample.

- Non-response : in many polls individuals some individuals do not respond. These individuals may be different in their behaviour, for example they may not "support" the current government but may not wish to say so.

- Wording : examples

  Do you support improved funding for education and health?

  Do you support increased tax to pay for education and health?

  In an opinion poll or survey the first question will typically obtain a higher support or yes level than the second question. This is often one of the difficulties for interpreting opinion poll surveys. Much effort is expended in careful psychological wording of opinion poll questionnaires.

Inference about the population

Random samples does eliminate systemic design bias in estimation when one does inference about the population.

This chapter discusses some aspects of obtaining data through sampling methods and designs. Well constructed designs can give valuable information.

This chapter does not discuss designed experimentation, the type typically performed in a laboratory or in scientific experimental. This later is performed to learn about causation, that is the effect if some treatment is given. These do not give information about how common this type of intervention is done in nature, so it cannot for example give information on the prevalence of some disease, but may give information on how it attacks an individual once the disease has arrived in its host.