Chapter 9 : Producing Data and Experiments

This chapter introduces the more ideas about how data is produced in a somewhat different manner than sampling (Simple Random Sample, Stratified Sample and Probability Proportional to Size). Here we consider data produced by experiments.

- objects or subjects : we wish to measure the outcome or value of this
- explanatory variable or factors
- treatment : refers to the condition applied to the subject

To help us understand these terms consider a typical medical experiment or clinical trial. Suppose we wish to determine if taking vitamin C is helpful in making colds less severe. In a clinical trial setting we would give one group of people either (i) a placebo (something with no effect) or (ii) a standard treatment. In another group we would give vitamin C, possibly in one of several doses. The people receiving either the placebo (or standard treatment) and those receiving the vitamin C are the subjects of the study. There are possibly several other explanatory variables also measured and recorded. These might be (i) categorical, such as state of health and gender, or (ii) numerical, such as weight, age, some physiological measurements. These would be used as explanatory variables, typically in some form of regression.

As with sampling methods, statistical techniques in the design of experiments are often essential if we wish to make useful inferences about the population who might be receiving the treatment, if is seems to be effective.

Example of bad design, or no design as likely would be the case in such a data collection. This is Example 9.3 of the 4th edition.

A graduate management admissions test (GMAT) preparatory course is offered by a university to students who are applying to their business school. In the past years only in class preparation is offered, but this year only the online version is offered. The grades of these students for the GMAT at 10% higher than in previous years. What does this mean in terms of *whether the online course is better or not*? Can one make any conclusion?

The subjects used for the test are typically people who are working in management or business and take the online course because it is possible to do so without leaving their current position or job. These students would have a very different motivation, experience, and are typically more mature than those taking the in class course. While these students have done better on the GMAT, it may be due to the online course, but it may be due to the factors described above. There is no way of determining this with the type of data collected. The *lurking* or hidden variables may be reason for the higher scores, and not the test preparation, or it may be the test preparation. This is the same question or idea of how do prove *causation* instead of just association. In sampling the method is randomization in choosing the subjects who are chosen to be in the sample. This removes *selection* bias, and a similar idea can be implemented in some forms of experimental data.

Experiments to Compare Treatments

Randomized comparative experiment

Subjects are assigned treatments at random.

Aside : There is an area of statistics called experimental design which studies ways of doing this with many treatments and in an efficient manner with as few subjects as possible.

Completely Randomized Design

Example 9.5 4th Edition, and how it can be used in Ontario which is installing SMART metres in 2009.

SMART metres are electricity use metres that are to help consumers know how much electricity they are using are various times during the course of a day. The *hope* is that electric power use will drop, at least during peak or high use times. Can one design an experiment to measure if SMART metres will work in this sense? There might also be more than one type of SMART metre.

A sample of households can be chosen in the province. For these households then a randomized experiment can be used. To be specific suppose there are two types of SMART metres under consideration. Thus there are 3 treatment groups, the standard electricity metre, SMART metre I and SMART metre II. The households in the sample can then be randomly assigned to be in one of the three *treatment* groups : (1) standard electricity metre, (2) SMART metre I and (3) SMART metre II. The electricity use will then be monitored over a period of time (how to choose this?) and then the results can be compared.

The data will be numerical values on the electricity consumption and we then compare three *treatment groups*, for example by graphical methods, some summary statistics, or a more formal method. We wish to examine a hypothesis that the three treatment groups are the same, versus an alternative that the treatment groups are not the same, and thus choose a best treatment.

Notice here that statistical methods can allow us to find the best *treatment* before installing huge numbers of SMART metres and then finding out after all this expense that a possibly poor choice has been made. Statistical methods will allow a good decision to be made before expensive commitments to some technology is made, thereby savings millions and perhaps billions of dollars.

The random assignment of subjects of individuals (in the example above households) removes selection bias which would remove any valid interpretation of causation.

Logic of randomized experiments

1. Random assignment of subjects to treatment groups will form groups that are similar in all respects *before* treatments are applied.
2. Comparative designs ensure that influences other than the experimental treatment operate equally on all groups. This is another way of saying that selection bias is removed.
3. The differences in response then *must be due* to either the treatments or random chance.

For the last point if we can understand or predict what range variables are highly likely to fall, then we can determine when something unusual happens. An *outlier* occurs when a very rare outcome happens by chance, or the mechanism to produce this outlier is different from that producing the rest of the data. This is an idea we pursued in Chapter 2.

For those who are interested in Sherlock Holmes, a famous fictional detective in stories by Arthur Conan Doyle, we are using something related to his deduction principle. If something is inconsistent with a null hypothesis (eg, SMART metres perform the same a regular metres), then the null hypothesis is not reasonable and so an alternative hypothesis must be true (eg SMART metres do not perform in the same fashion as regular metres).

Statistical Significance : An observed effect that is so large (so different) that it would rarely occur by chance is called statistically significant.

Principles of experimental design

1. Control the effects of lurking variables on the response, most simply by comparing two or more groups
2. Randomize the assignments of subjects to treatments
3. Use enough subjects (big enough sample sizes) to reduce the chance variation in the groups. This is often interpreted as reducing the variability in a useful summary statistic, such as sample means.

In many clinical trials where local physicians or hospital staff choose which patients get which treatment, there may still be selection bias, even if it is not a conscious bias. For this type of study and clinical trial, a *double blind* experimental design is used.

Example setting : To test the effectiveness of vaccine or some orally taken treatment, a placebo is often used as the control. A placebo is something which has no medicinal effect, that is it is neutral. In this setting these pills, and the treatment pill are made to look identical. They are typically put into an individual container with an identification number, so that the when the data is analyzed it will be know at that stage, but not to the physician, which treatment is given. A treatment is assigned at random to the subjects, using the unique identification number as a method to randomly select the treatment used. This is a double blind method. It is *blind* in the sense that the treatment assigned is at random, and again a second *blindness* in that the physician cannot use their bias about the health of the patient to modify the random choice of the treatment.

Human empathy might cause some physicians to otherwise change the treatment for some patients, thereby biasing the *population* characteristics of those who receive each type of treatment. If this were the case then inferences would be invalid as it would not be possible to determine if the effect observed is due to the treatment or to the differing population characteristics.

Matched pairs and other types of designs

Sometimes one can design experiments that are more effect than the simple randomized experimental designs.

Matched pair design : This is useful to compare two treatments. Pairs of subjects are chosen to match each other as closely as possible. For a given pair the 2 treatments are assigned randomly, one treatment to the first subject in the pair and the other to the second subject in the pair. The difference in the effects are then a measure of the difference in the treatment and removing the differences to common subject effect or subject bias.

Example : The matched pairs can be implemented in a very effective way in this example.

We wish to compare two car shock absorbers, say manufacturer A and B. This is typically done by installing them on a vehicle and then after some specified distance driven, say 50,000 km, the amount of force required to compress the shock absorber is then measured. What are the major factors that contribute to this final shock absorber strength, as measured here?

1. shock absorber A versus B
2. the weight of the car
3. the types of road it is driven on. One cannot control for all potholes and bumpiness of the road. Roads are also typically slanted from the centre towards the edges for drainage reasons.

The variability of the measurements, and hence of their sample means, are then effected by all these items. above. Can we reduce the variability of the sample means?

Suppose we put one of shock absorber A and one of B on each car, specifically on the back wheels. Then this pair of shock absorbers goes over the same road, and has the same weight of vehicle. Thus taking the difference of this pair removes the effects due to (2) weight of vehicle and (3) road type. Further randomizing whether shock absorber A

goes on the left or right side removes on average the tilt or the slant of the road. When we look at the average of the differences of the shock absorber strengths, some effects are then pairwise removed. Here a matched pair design is much more effective.

In some medical settings it is desirable to use identical twins, so that genetic differences for this pair is eliminated. For a long time it was observed that smoking and lung cancer were positively associated. Does this mean that smoking caused lung cancer? Might it not be that people who are genetically predisposed or highly acceptable to getting lung cancer might also be predisposed to desiring to smoke? This was the state of trying to interpret the observed smoking and lung cancer association in the first half of the twentieth century.

How was it finally concluded that smoking caused lung cancer? A statistician in the middle of the 20th century, R. A. Fisher, whose main work was in the design of experiments and genetics, suggested a design which involved studying smoking and lung cancer for pairs of identical twins. In this was one could rule out the *genetic predisposition* argument that was the controversy in attaching the causal conclusion that one ideally would like.

The same controversy is currently being played out in the concerns and interpretation of the observed global warming. What is causing this? Is it some natural cyclic variation of the earth and solar system? Is it an effect of man made atmospheric pollutants? Yes pollutants are put in the atmosphere, but is this coincidental with a natural warming period or not? These are often not trivial exercises to assign cause, no matter how evangelical one is for one side or the other.

Data and properly designed experiments can be very helpful. Mathematics and statistics (a special form of applied mathematics) are the language and tools that can be used.

When there are more than two treatments a block design may be used. With two blocks, there are 2 treatments (one pair of treatments).

- Block : group of individuals or experimental units that are similar or common
- Block design : treatments are assigned randomly and separately within each block (possibly with some constraint)
- the idea is to create treatment groups that are homogenous so that differences amongst these individuals is due to the treatment

Example : a type of hospital ward may have a few treatment types under study, for example heart attacks

Repeating this at several hospitals will have the effect of using hospital as a block.