3 Producing Data

In exploratory data analysis, we analyze data already collected and summarize them using graphs and numerical analyses. Sometimes, however we would like to obtain answers for a larger group of individuals and not just the individuals for whom we have data. To make sure answers from the data that we have can be extended to a larger group off individuals we need to make sure that the data is is collected in a way that is designed to answer our questions.

To answer the question "Is IQ and GPA related for high school students?" we would like to test for the relationship between IQ and GPA of all high school students, but it is not a feasible way to do it. However we could collect a subset of students of all possible high school students. This collection should be done in such a manner so that the subset represents all high school students. This subset is referred to a **"sample"** and the entire group of individuals we are interested in is referred to as **"population"**.

There are two ways to collect data from a sample.

- An **observational study** observes individuals and measures variables of interest, but does not attempt to influence the response.
- An **experiment** on the other hand, deliberately imposes some treatment on individuals in order to observe their responses.

An observational study is used when we would like to gauge the opinion of people or just measure some characteristic, but to understand the effects of some condition on an individual we need to conduct an experiment.

Consider for example when we would like to understand the influence of rainfall on yield and we try to collect data from farms in different cities. Note that we cannot be sure if rainfall or maybe something else like soil conditions or temperatures in the city had any effect on the yield. In this case an observational study cannot differentiate between the effects of rainfall, soil conditions and temperature and so the variables are said to be **"confounded"**, i.e. two variables are said to be confounded when their effects on a response variable cannot be distinguished from each other.

3.1 Designing Samples

A car manufacturer wants to know the opinion of people about improving a certain model of their cars. They want to get the opinion of people in North Carolina, but they are sure they cannot ask all their customers here because of time and cost considerations. In order to get information they ask a subset of their customers.

- The entire group of individuals that we want information about is called the **population**.
- A **sample** is part of the population that we actually examine in order to gather information.

In order for information from the sample to be useful, so that we can draw conclusions from it about the population, we need to make sure that the sample really represents the population. When web sites ask people to respond to a certain question, the people who respond can still form a sample but it really need not represent the population. The **design of a sample** indicates the technique which is used to construct the sample from the population.

When samples do not represent the population of interest but systematically favor some parts of the population over the others then these designs exhibit **bias** . Two sample designs which exhibit bias are

Voluntary Response Sample: is a sample where peole who choose themselves. When people are asked to express their opinion on a website or at a radio station, those are examples of voluntary response samples.

Convenience Sample: is a sample which is chooses the individuals easiest to reach. An example of a convenience sample would be when one sampled his class mates when he wanted information about GPA's of college students.

3.1.1 Simple Random Samples

In the above sampling designs (Voluntary response or Convenience), there is a personal choice involved (either of the people in the sample or the person who samples the population). This personal choice introduces bias in the results of the sample. To make sure that the selected sample is unbiased we need to make sure that there are no personal choices involved. One way to select such a sample is by chance. Giving all individuals in a population the chance of being in the sample reduces the bias.

The simplest way to select a sample using chance is by to place names of all individuals in the population in a hat and then draw out a handful. This is the idea of *simple random sampling* .

A **simple random sample (SRS)** of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.

How to choose an SRS?

Software: Instead of using a hat to get our random sample we can use computer software to generate our sample or we can use a calculator.

Random Digits: We can also randomize by using a table of random digits such as **Table B** . It is a long string of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 with these 2 properties

1. Each entry in the table is equally likely to be any of the 10 digits 0 through 9,
2. The entries are independent of each other. That is knowledge of one part of the table gives no information about any other part.

Choosing an SRS from Table B

Step 1: Label. Assign a numerical label to each individual in the population.

Step 2: Table. Use Table B to select labels at random.

An example on how to use the table!

An author, addicted to reading books decided to investigate the length of the books in his library, and find the average length. He does not have the time to look at all the books, so he decides to take a sample of size 5.

	0	1	2	3	4	5	6	7	8	9
0	229	247	347	246	307	181	198	214	234	340
1	314	260	202	320	360	320	200	414	262	248
2	376	211	214	218	276	628	255	352	197	308
3	203	371	203	406	261	378	223	181	284	196
4	278	167	329	213	373	521	133	312	685	255

The first step is to label all the population values. Since we have 50 books we can label them from 01-50 or from 00-49. The next step is to use Table B to select a sample of size 5. Select a line at random from the Table and based on the line choose the first five 2 digit values that are among the values of the labels that you have. If we choose line 111 as our line we have the following.

111	81486	69487	60513	09297	00412	71238	27649	39950

This line can be grouped as follows.

81 48 66 94 87 60 51 30 92 97 00 41 27 71 23 82 76 49 39 95 0

The first five numbers that satisfy our label conditions are 48, 30, 00, 41 and 27 (Assuming that we labeled the books 00-49). This would mean that the corresponding sample of books was 685, 203, 229, 167 and 352.

3.1.2 Other Sampling Designs

The general framework for statistical sampling is a **probability sample** which is defined to be a sample which gives each member of the population a known chance (greater than zero) of being selected in the sample.

We know that SRS gives each member of the population an equal chance of being selected. However that is not always the case. A **Stratified Random Sample** is one such case. Here the population is divided into groups of similar individuals, called *strata* . Then choose a seperate SRS in each stratum and combine them to form the full sample.

If you wanted to sample MLB pitchers, what are the possible strata?

Consider the example of the author. Suppose each row was a shelf of similar type of books. If each row was a stratum construct a sample of size 5 taking one book from each row!

3.1.3 Multistage Samples

One of the most popular sampling techniques is multi-stage sampling where sampling is done at many levels. **Cluster Sampling** is a sampling design where the population is divided into clusters. Choose a random sample of clusters and then choose a sample from each cluster selected. This is an example of two-stage sampling.

3.1.4 Cautions about sample surveys

There are many problems when conducting a sample survey. Bias in sample selection is one of the main issues, but random selection takes care of bias. There are other issues which however might bias the data.

Undercoverage: Sometimes our population list is not complete and so some parts of the population are not represented in the sample. This introduces bias in the results of the survey.

Nonresponse: occurs when an individual chosen for the sample cannot be contacted or refuses to cooperate.

Another way of introducing bias is through the questions of the survey. Questions which require memory to be good or sensitive questions may not elicit the right answer. This kind of bias is known as **response bias** . Also some questions can be misleading or confusing beacuse of the **wording (wording effects).** These questions introduce more bias.

What kind of bias?

1. How often did you eat out in the last fortnight?
2. How often did you have sex in the last month?
3. Should laws be passed to eliminate all possibilities of special interests giving huge sums of money to candidates?
4. Should laws be passed to prohibit interest groups from contributing to campaigns, or do groups have a right to contribute to the candidates they support?

3.1.5 Inferences about the population

Since we cannot get information from the population we can never be sure of the accuracy of the results form the sample. It is more likely that two samples will give two different conclusions about the population than the same. So under such circumstances, how much can you trust the results of your sample survey?

If the sampling technique was based on random sampling, then the laws of probability govern it and we can get error bounds for our answers. Also by increasing sample sizes we can get more accurate results.

3.2 Designing Experiments

As mentioned before, sample surveys give information on some characteristic, but cannot distinguish between the effects of two characteristics.

To understand the effects of one variable on another we need to conduct an experiment. Here is the vocabulary of experiments

Experimental Units: The individuals on which the experiment is conducted.

Subjects: When the experimental units are humans, we refer to them as subjects.

Treatment: A specific experimental condition applied to the units is called a treatment.

Factor: The explanatory variables in an experiment are often called factors.

Level: is a specific value of one or more factors, the combination at which form the treatment.

Consider a study on the effect of two drugs A and B at 3 levels (high, medium and low dosage) on the reduction of blood pressure. In this example the **subject** are the humans undergoing the test, the **factors** are drugs A and B, the **levels** are 1, 2 and 3 and the **treatments** are the combination of the dosages. To be precise there are 6 treatments (High A and High B, High A and Medium B, High A and Low B and so on...)

Consider the effect of rainfall on yield. If we wanted to test the effect of rainfall on yield we might conduct the experiment in a green house and vary rainfall types. The **experimental units** are the plants, the **factor** is rainfall and the **levels** are the different levels of rainfall. By changing rainfall levels and measuring yield we can find out the effect of rainfall and yield.

However we have to make sure that other factors are held constant. We can change rainfall, but if we change temperature also, then we cannot be sure if the change in yield is due to the change in rainfall levels or the change in temperature levels. We need to keep temperature constant so that we can attribute any change in yield to change in rainfall.

3.2.1 Comparative Experiment

Experiments typically have a simple design.

Units → Treatment → Response

i.e. each unit is subjected to the treatment and the response is measured. However we have to make sure that any changes in respnse is due to the treatment and no other factor.

Gastric Freezing was a clever method to treat ulcers. The patient swallows a deflated balloon with tubes attached and a refrigerated liquid is pumped through the balloon for an hour. The idea is that cooling the stomach will reduce the its production of acid and so relieve ulcers. An experiment showed that gastric freezing did reduce acid production and relieve ulcer pain. However the patients were responding to what is called a **placebo effect.**A placebo is a dummy treatment, mainly used to weed out the factor that patients sometimes respond to any treatment. In this case the patients were responding to the fact that they were being treated. The treatment for ulcers has since be dropped. The placebo effect is a **lurking variable** and as such we need to control its effects by using a **control group** which receives a dummy treatment.

3.2.2 Randomized comparative experiments

A design of an experiment describes the response variables, the factors (explanatory variables), and the layout of the treatments, with comparison as the leading principle.

The second aspect of the design is the rule used to assign the experimental units to the treatments. How to assign individual units to treatment combinations? We use chance to assign experimental units to treatments and this is referred to as **randomization.**

3.2.3 Completely randomized designs

A **completely randomized design** is an example of a randomized comparative experiment. In this design all the experimental units are allocated at random among all treatments.

Many utility companies have introduced programs to encourage energy conservation among their customers. An electric company is considering placing electronic indicators in households to show what the cost would be if the electricity use at the moment continued for a month. Would cheaper methods be as effective? The alternative to indicators is to use a chart and information about monitoring their electricity use. Design an experiment for the company if they have 60 household willing to participate.

3.2.4 The logic of randomized comparative experiments

The logic behind the use of experimentation is to get evidence that differences in treatments cause changes in response. Randomized Comparative designs help us acheive that because of the following.

- Random assignment forms groups that should be similar in all respects before the treatments are applied.
- Comparative design ensures that influences other than the experimental treatments operate equally on all groups.
- Hence, differences in average response must be due to the treatments or to the play of chance in the random assignment of experimental units to treatments.

The last point is an important one. In the electricity comapny example by random chance we could end up with 20 households with bad insulation ending up in the control group and that could be a factor in the variation of responses. However if we were to repeat the exeriment a few times, it is highly unlikely that that assignment would repeat itself. So this leads us to the three priniciples of experimental design.

1.Control of the effects of lurking variables on the response, most simply by comparing several treatments.

2.Randomization, the use of impersonal chance to assign experimantal units to treatments.

3.Replication of the experiment on many units to reduce chance variation in the results.

Sometimes the difference in the responses is so large that it is very unlikely that it could have happened just by chance. Such effects are referred to as being **statistically significant.**

3.2.5 Cautions about experimentation

In all experiments, one should treat all experimental units equal, until the treatments themselves are applied. However knowledge of the treatment being applied might change the approach of the one conducting the experiment. To take care of this issue sometimes experiments are **double-blind** . Double-blind experiments are those where neither the subject nor the personnel who work with them knew which treatments any subject had received.

In experiments one of the primary issue is the ability to replicate the conditions we want to actually study. Many experiments suffer to this **lack of realism** issue. When someone knows they are being studied, we can never be sure if their reactions are the same as one who does not know he is being observed.

3.2.6 Matched pair design

When you have to compare two treatments, it is advisable to give each subject or experimental unit both the treatments so one can compare the effects of both the treatments everything else being held the same. Note that it cannot be done all the times because we might have treatments which have a residual effect or when we want to test two drugs to cure a life-threatening disease. However if we wanted to compare two pens for ease of writing we could have each subject use both the pens and compare the two. This type of design is called a matched pair design.

3.2.7 Block designs

In a CRD, we randomly assign treatments to experimental units. However like in the matched pair design we can do better by not doing so. We only randomize within each matched pair. This is an example of a block design.

A **block** is a group of experimental units or subjects that are known to be similar in some way that might affect the response to the treatments. In a **block design** , the random assignment of units to treatments is carried out seperately within each block.

Blocking helps by removing any variation due to the similarities of certain units and their difference from the other units. It also allows us to draw seperate conclusions for each block and gives more precise conclusions about response changes than a CRD.