Topic 4
Producing Data
Formal Statistical Procedures

Outline

Observational Studies

The goal is of an observation study is to learn about a population by observing a sample with as little disturbance as possible to the sample.

- Sometimes the selection of treatments is not under the control of the researcher.
  - If we suspect that a certain mutation would render a virus more or less virulent, we cannot ethically perform the genetic engineering and infect humans with the viral strains.
- Effects are often confounded and thus causation is difficult to assert.
  - The link between smoking and a variety of diseases is one very well known example. We can see that children of smokers are more likely to smoke. This is more easily described if we look at conditional distributions

Randomized Controlled Experiments

- In a randomized controlled experiment, the researcher imposes a treatment on the experimental units or subjects in order to observe a response.
- A good experimental design is one that is based on a solid understanding of both the science behind the study and the probabilistic tools that will lead to the inferential techniques used for the study.
  This study is often set
  - to assess some hypothesis - *Do parents smoking habits influence their children?* or
  - to estimate some value - *What is the mean length of a given strain of bacteria?*

Principles of Experimental Design

1. Control for the effects of lurking variables by comparing several treatments.
2. Randomize the assignment of subjects to treatments to eliminate bias due to systematic differences among categories.
3. Replicate the experiment on many subjects to reduce the impact of chance variation on the results.

Issues with Control

The desired control can sometimes be quite difficult to achieve. For example;

- In medical trials, some may display a placebo effect.
- Overlooking or introducing a lurking variable can introduce a hidden bias.
- The time and money invested can lead to a subconscious effect by the experimenter. Use an appropriate blind or double blind procedure.
- Changes in the wording of questions can lead to different outcomes.
- Transferring discoveries from the laboratory to a genuine living situation can be difficult to make.
- The data may suffer from undercoverage of difficult to find groups.
- Some individuals leave the experimental group, especially in longitudinal studies.
- In some instances, a control is not possible.
- Some subjects may lie.

Setting a Design

Before data are collected, we must consider some basic questions:

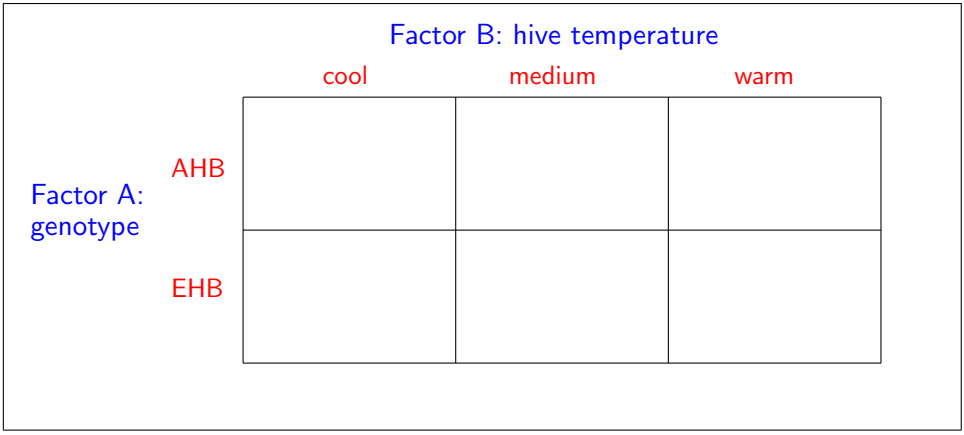
- Decide on the number of explanatory variables or factors.
- Decide on the values or levels that will be used in the treatment.

Example. A breeding experiment using African and European bees occurred in 1956 in an apiary in the southeast of Brazil. The hybrid bees escaped and today, in the western hemisphere, all Africanized honey bees are descended from this apiary.

When the time arrives for replacing the mother queen, she will lay about 10 queen eggs. Let's investigate the question of whether a shorter time for development for Africanized bee queens is the mechanism behind the replacement by the Africanized subspecies. Development depends upon hive temperature. We will set three levels - cool, medium, and warm. European honey bee queens are the control.

Setting a Design

Thus, this experiment has 6 treatment groups.



Random Samples

A simple random sample (SRS) of size *n* consists of *n* individuals chosen in such a way that *every* set of *n* individuals has an *equal chance* to be in the sample actually selected. This is easy to accomplish in R. using the command `sample` . For the experiment above, we rear 90 Africanized queens and choose a sample of 60.

```
> population<-c(1:90)
> (subjects<-sample(population,60))
 [1] 61 16 65 73 13 25 10 82 24 62 28 66 55  8 26 72 67 17 58 69  6 27 41 20
[25] 87 68 22 11  5 48 33 63 50 88 35 37 84 12  4 59 90 86  2 60 19 18 74 23
[49] 78 49 45  7 64  3 42 57 81 56 46 32
```

Random Samples

For experimental designs that call for grouping similar individuals, called strata, then a stratified random sample from the full sample by choosing a separate random sample from each stratum. A stratified random sample ensures the desired number of sample from these groups is included in the sample.

If we mark the 180 queens 1 through 180 with 1 through 90 being Africanized bees and 91 through 180 being European, then we can enter

```
> population<-c(1:180)
> subjectsAHB<-sample(population[1:90],60)
> subjectsEHB<-sample(population[91:180],60)
```

to ensure that 60 come from each group

Exercise

Exercise. A health study is being conducted on a population of 20 women and 40 men.

1. Label the women 1 through 20 and the men 21 through 60. Estimate how many women are in a simple random sample of size 24.
2. Determine 10 simple random samples and record the number of women in each of the samples.
3. Find the mean and the standard deviation of the number of women in the samples. How does this conform to your estimate?
4. Perform a stratified random sample having 12 women and 12 men and report your findings.

Case Study

Example. Salk vaccine field trials

- Poliomyelitis is an acute viral infectious disease spread from person to person, primarily via the fecal-oral route.
- The overwhelming majority of polio infections have no symptoms. However, if the virus enters the central nervous system, it can infect motor neurons, leading to symptoms ranging from muscle weakness and paralysis.
- The effects of polio have been known since prehistory. However, the first US epidemic was in 1916. By 1950, polio had claimed hundreds of thousands of victims, mostly children.

Case Study

In 1950, the Public Health Service (PHS) organized a field trial of a vaccine developed by Jonas Salk. Polio is an epidemic disease with

- 60,000 cases in 1952, and
- 30,000 cases in 1953.

So, a low incidence without control could mean

- the vaccine *works*, or
- *no* epidemic in 1954.

Case Study

Some basic facts were known before the trial started:

- Higher income parents are more likely to consent to allow children to take the vaccine.
- Children of lower income parents are thought to be less susceptible to polio. The reasoning is that these children live in less hygienic surroundings and so are more likely to contract very mild polio and consequently more likely to have polio antibodies.

Case Study

At the same time as the PHS study, a parents advocacy group, the National Foundation for Infantile Paralysis (NFIP) set out its own design. Here are the essential features of the NFIP design:

- Vaccinate all grade 2 children with parental consent.
- Use grades 1 and 3 as controls.

This design fails to have some of essential features of the principles of experimental design. Here is a critique:

- Polio spreads through contact, so infection of one child in a class can spread to the classmates.
- The treatment group is biased towards higher income.

Case Study

The Public Health Service design is intended to take into account these shortcomings. Their design has the following features:

- Flip a coin for each child. (randomized control)
- Children in the control group were given an injection of salt water. (placebo)
- Diagnosticians were not told whether a child was in treatment or control group. (double blind)

The results:

	PHS		NFIP	
	Size	Rate	Size	Rate
Treatment	200,000	28	225,000	25
Control	200,000	71	725,000	54
No consent	350,000	46	125,000	44

Rates are per 100,000