# Introduction to Data Collection

UBCO MDS — DATA 543

# Reference Material

This module was originally developed by Dr. Irene Vrbik with later contributions by Dr. Jeff Andrews.

We will be pulling material from multiple sources including:


**Lawson (2014)** John Lawson. *Design and Analysis of Experiments with R*. Chapman and Hall/CRC, 2014.

**Dean (1999)** Angela Dean, Daniel Voss, and Danel Draguljić. *Design and Analysis of Experiments, Second Edition*.

**Montgomery (2017)** Montgomery, Douglas C. *Design and Analysis of Experiments, 8th Edition*.

**Tamhane (2012)** Ajit C. Tamhane. *Statistical Analysis of Designed Experiments: Theory and Applications*.

**Lohr (2009)** Lohr, Sharon. *Sampling: design and analysis*.


**Quiz1**

# Introduction

*We are drowning in information but starved for knowledge. Uncontrolled and unorganized information is no longer a resource in an information society, instead it becomes the enemy.*

John Naisbitt, Megatrends (1982)

# Introduction

- **Data collection** involves gathering information on variables of interest. Data is usually collected through sampling surveys, observational studies, or experiments.

- The collected variables are commonly motivated by some **research question**.
  - What is the average age of students in the MDS program at UBCO?
  - Does taking vitamin C decrease the duration of a cold?

# Types of Research Questions

**Descriptive:** attempts to explore and explain characteristics of a population/phenomenon. For example,

- who is most commonly buying this product?
- frequencies, averages, and other statistical calculations

**Exploratory:** initial research which investigates a hypothesis or theoretical idea. Lays the groundwork for future research.

**Causal:** attempts to make a connection between a cause (say process 1) with an effect (say, process 2) where process 1 is partly responsible for process 2, and process 2 is partly dependent on process 1.

- will sales increase with a change in X?

# Descriptive vs. causal (inference)

Causal (or inferential) research is a much harder task. For example,

- if we discover that our cold only lasted 4 days while taking vitamin C, there is no way of knowing if our cold would have been longer/shorter had we not taken vitamin C.
- if we discover a relationship between X and Y, did X cause Y, did Y cause X or did Z cause both?
- The validity of any scientific results depends on accurate and honest measurements.
- Collecting data using systematic, established techniques is a critical first step in data analysis.

# Types of Data collection

**Sampling surveys** are useful to estimate some property/parameter of a finite population without conducting a census

- broadly, a *census* is a list of all individuals in the population along with characteristics regarding each individual.
- Eg. the proportion of registered voters in a particular precinct that favour a proposal could be estimated by polling a random sample of voters rather than questioning every registered voter in the precinct.

**Observational studies and experiments** are used to determine the relationship between *two or more* measured quantities.

- Eg. the relationship between blood pressure and screen time.

# Observational studies vs. experiments

- *Observational studies* passively observe data in its natural environment and attempts to draw conclusions about associations between them.

- Observational studies can only show *association* (eg. correlation between two variables), not *causation* (one variable produces an effect on another variable).

- *Why not?* Both observed variables may be affected by changes in a third variable (either *confounding* or *lurking* variable) that was not observed or recorded.

# Observational studies vs. experiments

Take this example from pg 2 of Tamhane (2012):

> *An observational study may find that people who exercise regularly live healthier lives. But ... there are many other variables such as diet, sleep, and use of medication that can affect a person's health. People who exercise regularly are likely to be more disciplined in their dietary and sleep habits and hence may be healthy. These variables are not controlled in an observational study and hence may confound the outcome. Only a controlled experiment in which people are randomly assigned to different exercise regimens can establish the effect of exercise on health.*

# Observational Studies: Example

Source: pg 2 Dean (1999)

Suppose the output from machines in a factory finds that a particular machine is consistently of low quality.

- Does the machine need replacing? Is the machine operator at fault? Is the humidity in that part of the factory suboptimal?
- The observational study shows that a problem exists but does not readily tell us the cause of the poor quality.

These variables are not controlled in an observational study and hence may confound the outcome.

# Observation vs. experimental

- In an *experiment* the environment is controlled so that some variables are purposely changed while others are held constant.
- Causation can only be established in a controlled experiment in which variable(s) is directly manipulated to see the influence on another variable.

# Experiment: Example

Source: pg 2 Dean (1999)

- Let's consider controlling for the operator from the example from slide 10.
- Our experiment moves all the operators from machine to machine over several days
- If the poor output follows the operator, then it is safe to conclude that the operator is the cause.
- If the poor output remains with the original machine, then the operator is blameless.

Here we have controlled over *one* possible cause in the difference in output quality between machines. If this particular cause is ruled out, then the experimenter can begin to vary other factors such as humidity or machine settings.

# Experiments

General goal: to better understand the phenomenon.

- eg., *what* are the key factors affecting the outcome of this phenomenon.
- *how* are the key factors affecting the outcome of this phenomenon.

This process is often iterative or sequential.

# Comment

- While the same methods may be applied on observational studies and experiments, stronger conclusions are possible from experiments.

- Experiments are the basis of the scientific method and are applied in countless fields

# Definitions

But first some basic definitions from Lawson (2014)...

- An *experiment* (aka a *run*) is an action where the experimenter purposefully changes at least one of the variables being studied and then observes the effect of his or her action(s).
- A *factor* (aka *independent variable* or *treatment factor*) is a controlled independent variable whose *levels* are set by the experimenter.
- The different settings of a treatment factor are called its *levels*. The levels change systematically from run to run in order to determine what effect it has on the response(s).
- a *treatment* is something that researchers administer to experimental units, i.e a particular combination of factor levels.

| | | **Factor 1** | | |
|---|---|---|---|---|
| | | *Level 1* | *Level 2* | *Level 3* |
| **Factor 2** | *Level 1* | Trt 1 | Trt 2 | Trt 3 |
| | *Level 2* | Trt 4 | Trt 5 | Trt 6 |
| | *Level 3* | Trt 7 | Trt 8 | Trt 9 |
| | *Level 4* | Trt 10 | Trt 11 | Trt 12 |

# Definitions

- an *experimental unit* is the physical entity (e.g., patients, plots of land, items) which receives a particular treatment and whose responses are then observed.

- a *replication* is the number of independent instances of a treatment that occur within an experiment i.e. several experimental units receive the same treatment.

- All experimental units receiving the same treatment form a *treatment group*.

# Definitions

- A *dependent variable* (or the *response* denoted by $Y$) is the characteristic of the experimental unit that is measured after each experiment or run.

- The *effect* is the change in the response that is caused by a change in a factor or independent variable.

- The *experimental design* is a collection of experiments or runs that is planned in advance of the actual execution. The particular runs selected in an experimental design will depend upon the purpose of the design.

It is important to note that experimental units (EU) must be independent.

- For example, if a batch of buns are made according to recipe A (treatment), the batch—not the individual buns—is the EU.
- Note that a replicate could be obtained by baking another batch using the same recipe.

Another important component is that the allocation of a treatment to a particular experimental unit is random.

- Consider testing whether a new strain of potato produces a higher crop than the standard one. . .

# Definitions

In contrast to an independent replicate, a *repeat measurement* uses the same EU to measure multiple responses.

- Ex: measurements (eg. internal temperature, amount risen) taken on buns from the same batch are repeated measurements.
- Longitudinal studies; repeated measurements are collected on the same patient

*Duplicates* refer to duplicate measurements of the same experimental unit from one run or experiment.

- The measured dependent variable may vary among duplicates due to measurement error
- in the analysis of data these duplicate measurements should be averaged and not treated as separate responses.

# Observation vs. experimental

More formally, . . .

***Observational studies*** record **factors** or **independent/explanatory variables** and **dependent/response variables** and draws conclusions about associations between them.

***Experiments*** actively manipulate the **factors** and evaluates their effects on the **response variables**.

- Because we set the levels of the treatment factors, they are said to be *controllable factors*.
- Of course in any experiment there will always be factors that we do not or can not control for.

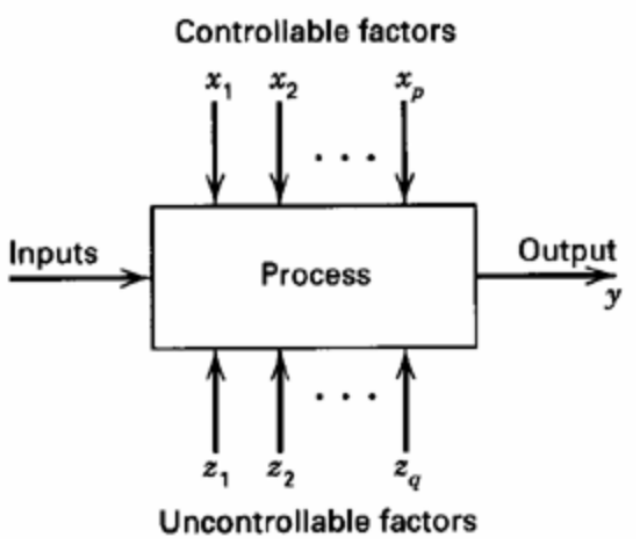


**Figure 1-1** General model of a process or system.

Image Source: Montgomery (2017)

- A *lurking* (aka *background*) variable is a variable that the experimenter is unaware of or cannot control.

- *Confounded variables* arise when each change an experimenter makes for one factor, between runs, is coupled with an identical change to another factor. In this situation it is impossible to determine which factor causes any observed changes in the response or dependent variable.

- **NB** (Nota bene meaning "mark well") the terms confounding and lurking variables are often used interchangeably.

# Experiment example: Box et al. (1978)

Consider an experiment[1] whose purpose was to determine whether a change in the fertilizer mixture would result in a change in the yield of tomato plants. Eleven tomato plants were planted in a single row, and the fertilizer type (A or B) was varied

**Treatment factor:** the type of fertilizer applied.

**Experimental unit:** the tomato plant plus the soil it is planted in

**Response variable:** hmmm...we'll come back to this

[1] see pg 10 in Lawson (2014)

# Experiment example: Box et al. (1978)

The possible conclusions the experimenter could draw:

(1) deciding that the fertilizer has no effect on the yield of tomatoes

$\Longrightarrow$ choose to use the less expensive fertilizer

(2) concluding that one fertilizer produces a greater yield.

$\Longrightarrow$ raises the questions:

Q: does the increase in yield offset any increase in cost of the better fertilizer?

Q: how large a difference in yield should he look for?

Another question crucial in planning the experiment is: *how many tomato plants should he include in his study?*

# Experiment example: Box et al. (1978)

The experimenter should consider the similarity or homogeneity of plants (i.e. EU) and how far apart he is going to place the tomato plants in the ground.

- Will it be far enough that the fertilizer applied to one plant does not bleed over and affect its neighbours?

# Experiment example: Box et al. (1978)

- The response variable is tricky in this experiment.
- Consider the weight of tomatoes: not all the tomatoes on a single plant ripen at the same time.
- Is it the weight of all tomatoes on the plant at a certain date, or the cumulative weight of tomatoes picked over time as they ripen?

Precision in the definition of the response and consistency in adherence to the definition when making the measurements are crucial.

# Experiment example: Box et al. (1978)

- There are many possible lurking variables to consider in this experiment.
- Any differences in watering, weeding, insect treatment, the method and timing of fertilizer application, and the amount of fertilizer applied may certainly affect the yield
- The experimenter must pay careful attention to these variables to prevent bias.
- The row position seems to have affected the yield as well (see Figure on next slide).

Figure 1.2 *Plot of Yield by Row Position—Tomato Experiment*

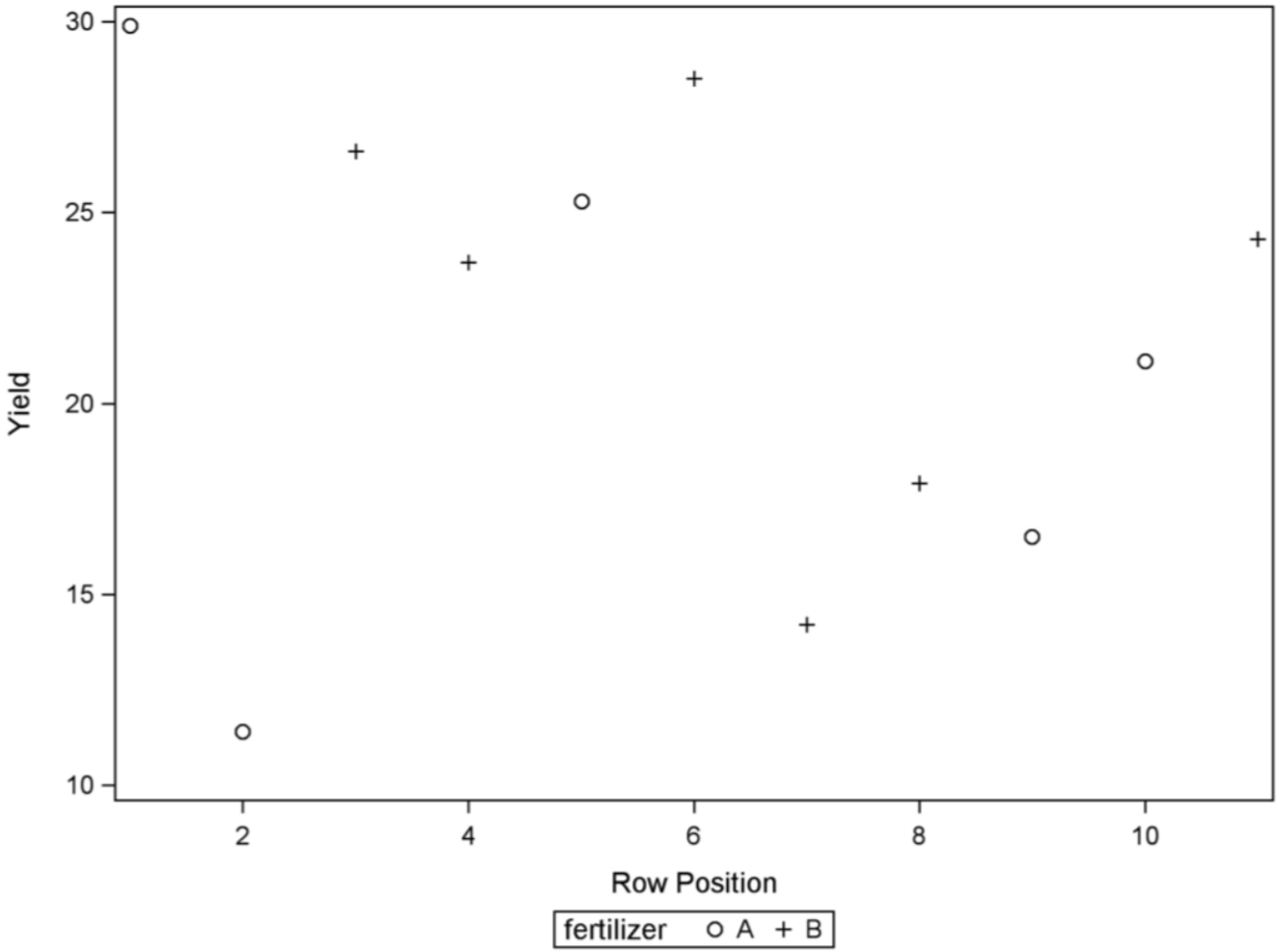


Image Source: Lawson (2014)

# Experiment example: Box et al. (1978)

- *Randomizing* which plants get which fertilizer works towards preventing systematic and personal bias from being introduced.
- Suppose the first 6 plants were given fertilizer A, and the remaining, B instead of random assignment. Then the row position effect could have been mistaken for a treatment effect!
- To control for the row position effect the adjacent pairs of plots could have been grouped together in pairs, and one fertilizer assigned at random to one plot-plant in each pair.
  - This technique is called blocking and will be discussed in more detail later.

# Experiment example: Box et al. (1978)

- One question raised by Easterling (2004) is *why were only eleven plants used in the study* (normally tomato plants come in flats of twelve).
- It could be, that one plant was removed from the study because it appeared unhealthy or got damaged in handling.
- Could this also explain why the yield for the plant in row position 2 was considerably lower than the others planted in neighbouring row positions with the same fertilizer? i.e. was this plant unhealthy or damaged as well?

This module will discuss two important areas in statistics:

(1) survey sampling; and

(2) design and analysis of experiments.

Survey sampling typically refers to observational studies.

- Study a *sample* from a *population*.
- Passively observe and record the variable(s) of interest.

Experimental design refers to conducting experiments.

- Actively manipulate the factors and evaluate their effects on the response variables.
- A population is (at least conceptually) infinite.

# Survey Sampling

## Why Sample?

We conduct samples to learn about a **population**. e.g.

- What is the month-by-month unemployment rate in Canada?
- What proportion of UBC students eat on campus?

Key aspect: our population is **finite**.

In theory, we could conduct a **census** and survey everyone of interest, but this will often be impractical.

# Experimental Design

There are many types of experimental designs. The appropriate one to use depends upon the objectives of the experimentation.

While there are many factors that go in to designing a good experiment, the three basic principles of experimental design are:

- Randomization
- Replication
- Reducing Noise (blocking)

i.e. the "Three Rs"

# Randomization

*Randomization* is the random allocation of treatments to the experimental units.

**The antityphoid inoculation example**:

> *Sir Almroth Wright, a famous immunologist, compared the incidence of typhoid between those who volunteer and those who refused the inoculation. Karl Pearson (the other Father of Statistics) noted that a volunteer may be more particular about their own health, thus more likely to run a lower risk.*

The above example, does *not* randomize properly. Experiments that properly randomize assist in "averaging out" the effects of extraneous factors that may be present.

# Replication

*Replication*: the repetition of a treatment within an experiment.

- Repetition of experimental conditions increases the accuracy of estimates which allows for estimates of the natural variation between experimental units.
- Knowledge of this variation will help generalize any relevant conclusions to similar subjects.

# Reducing Noise (Blocking)

*Reduce noise*: control the conditions in the experiment as much as possible

- Blocking can often be used to control and adjust for some of the variation in experimental units.
- *Blocking* divides experimental units into subsets (blocks) so that units within the same block are more similar than units from different subsets or blocks.

# What's to come

- Next class, we will begin a more detailed exploration of survey sampling.

- The second half of the course will focus on experimental design.

# References I


Dean, A., V. D. D. D. e. a. (1999), *Design and analysis of experiments*, Vol. 1.

Lawson, J. (2014), *Design and Analysis of Experiments with R*, Vol. 115.

Lohr, S. (2009), *Sampling: design and analysis*, Nelson Education.

Montgomery, D. (2017), *Design and analysis of experiments*, John Wiley and sons.

Tamhane, A. (2012), *Statistical Analysis of Designed Experiments: Theory and Applications*, Wiley Series in Probability and Statistics, Wiley.
**URL:** *https://books.google.ca/books?id=iKokB7x_31cC*