Introduction & Definitions

Sean Hellingman ©

Design for Data Science (ADSC2030)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- Recall: *In the field of statistics, we collect, analyze, and draw conclusions from data.*

- Sampling surveys are often used to estimate some property of a *finite* population.
  - Example: The proportion of TRU students that enjoy data science.

- Observational studies and experiments are generally used to determine the relationship between two or more measured quantities in a *conceptual* population.
  - Example: Interested in the relationship between future greenhouse gas emissions and future average global temperature (population does not yet exist).

Observational Studies & Experiments

- In an **observational study**, data is observed in its natural environment.
  - It *cannot* be proven that the relationships detected are cause and effect.
- In an **experiment** the environment is controlled. Some variables are purposely changed while others are held constant.

Planned Experiments

- Experiments may be used to:
  - Determine the cause of variation in some response variable.
  - Find conditions that create the maximum or minimum response.
  - Compare the response between different settings of controllable variables.
  - Predict future response values.

- Presently, experiments are used across many fields including:
  - Engineering design
  - Industrial research
  - Biological science
  - Business management
  - Psychology

Definitions

Experiment

- An **experiment (run)** is an action where the experimenter changes at least one of the variables being studied and then observes the effect of this change.

- General collection of *observational* data is **not** an experiment.

Experimental Unit

- An **experimental unit** is the item under study upon which something is changed.
- When the experimental unit is a person the units are called **subjects**.

Sub-Sample

- A **Sub-Sample/Sub-Unit/Observational Unit** occurs when the experimental unit is split, *after* the action has been taken upon it.

- *Measurements on sub-samples, or sub-units of the same experimental unit, are usually correlated and should be averaged before analysis of data rather than being treated as independent outcomes.*

Independent Variable

- An **Independent Variable/Factor/Treatment Factor** is one of the variables that is being controlled.
- The *level* is being changed in a systematic way for each run to determine the impact of the changes on the *response(s)*.

Background Variable

- A **Background Variable/Lurking Variable** is a variable that cannot be controlled for, or the experimenter is unaware of.
- If the design is well planned, the effects of Lurking variables should balance and not impact the conclusions of the study.

Dependent Variable

- The **Dependent Variable/Response Variable** is the characteristic of the *experimental unit* that is measured after each experiment or run.
- The values of the response depends upon the settings of the independent variables or factors and lurking variables.

Effect

- The **Effect** is the change in the response caused by a change in a factor or independent variable.
- The effect is estimated through the observed response data after the *runs* in the experimental design have been conducted.
- The *practical effect* or the *size of a practical effect* is how large the effect should be to have practical importance (may be known prior).

Replicate

- **Replicate** runs are two or more experiments conducted using the same combinations of factors or independent variables on different experimental units.

- Used to minimize the potential impact of any lurking variables or natural differences in experimental units.

Duplicates

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

- Should NOT be treated as separate responses and the only variability should be due to measurement error.

Experimental Design

- **Experimental Design** is a collection of experiments or runs that are planned in advance.
- The particular runs selected will depend upon the purpose of the design.

Confounded Factors

- **Confounded Factors** occur when each change to a factor, between runs, is paired with an identical change to another factor.
- When this occurs, it is impossible to determine with factor causes any observed changes in the dependant variable.

Biased Factor

- A **Biased Factor** occurs when changes are made to an independent variable at the same time when changes in background or lurking variables occur.

- When this occurs, it is impossible to determine if any changes to the dependant variable are due to changes in the factor or changes in other background/lurking variables.

Experimental Error

- **Experimental Error** is the difference between the observed response for a specific experiment and the long run average of the results of all experiments conducted using the exact same combination of variables and factors.

- There are broadly two types of errors:
  1. Bias error: tends to remain constant or change in a consistent pattern over the runs in an experimental design.
  2. Random error: changes from one experiment to another in an unpredictable manner and average to be zero.

Observational Studies & Experiments II

- In an **observational study**, variables are observed without any attempt to change/control independent factors.
  - Background or lurking variables *may* be the cause of any changes in the dependent/response variable.

- In an **experiment** the independent variables are purposely controlled and varied and runs are executed in a way to minimize the impact of any lurking variables.

Purposes

Purposes of Design I

- Using experimental designs gives a recipe for a successful application of the scientific method.

- Iterative process with the following steps:
  1. Observing *nature* (something of interest).
  2. Hypothesizing some mechanism for what has occurred.
  3. **Collecting data**.
  4. Analysis to confirm or reject the hypothesis.

- Experimental designs provide a clear plan for data collection based on the hypothesis.

Purposes of Design II

- Using experimental designs allows the user to avoid the effects of confounding factors.
- When we conduct physical experiments, the response variable will differ over replicate runs due to differences in experimental units.
  - We use our designs to minimize this *experimental error*.
  - **Error control methods:** randomization, replication, and blocking.
- May also be designed to capture the magnitude of the impacts of factors.
- **The objective of a research program dictates which type of experimental design should be utilized.**

Illustrative Example 1

- **Randomization** helps to avoid confusion or biases due to changes in background or lurking variables.

- In 1954 a trial of a polio vaccine was tested on over 1.8 million children.

- The initial plan was to only give students in grade 2 the vaccine and grades 1 and 3 the placebo.
  - Rejected because the doctors would know who received the vaccine and it would bias their diagnosis. (similar symptoms)

  - In this plan the factor purposely varied, vaccinated or not, would have been biased by the lurking variable of doctors' knowledge of the treatment.

  - Instead, a double-blind study was used.

Types of Design I

- There are many types of experimental designs.
  - Selection depends on the objective of the experimentation.
- Two main objectives categories:
  1. To study the source(s) of variability in a response variable.
  2. To establish cause and effect relationships between variables.
- Experiments to study the source of variability are often a starting point before examining cause and effect.
  - Which independent variables should be considered for the cause and effect analysis.

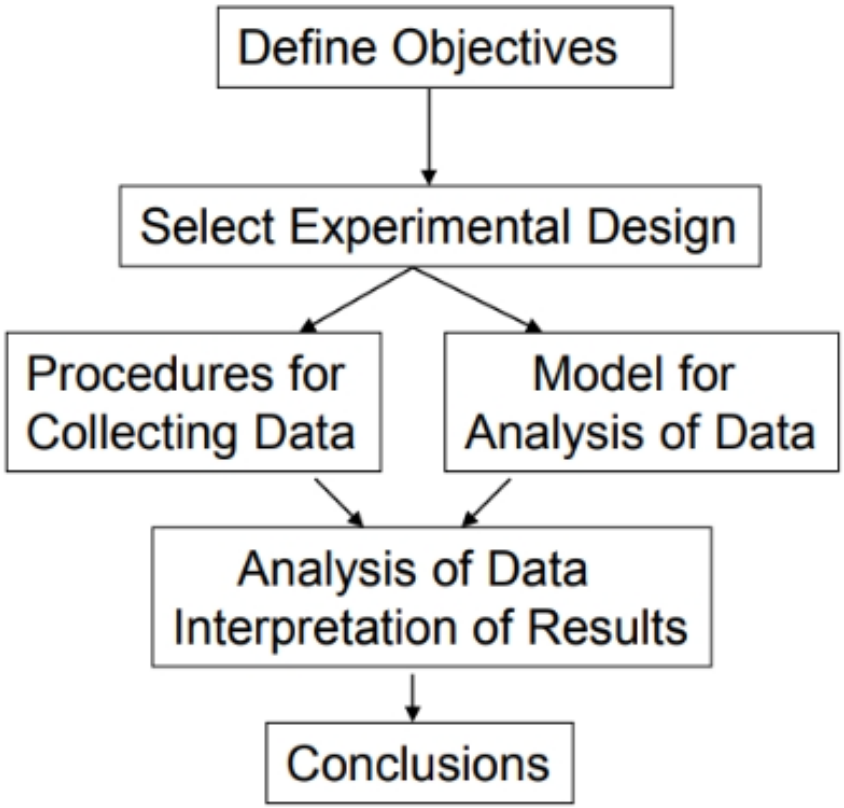
Types of Design II

- Initial steps involve identifying the important treatment factors, their effects are quantified, and the *optimal* combination can be determined.

- The most important factors are identified during the **screening stage.**
  - Steps taken early in the process to uncover the most important factors.

- Choose between:
  1. **Constrained optimization:** usually six or fewer factors under study.
     - Used to quantify the effects of the factors, their interactions, and identify the optimal combination.
  2. **Unconstrained optimization:** Trying to map the relationship between one or more responses and five or fewer quantitative factors.
     - Used to improved operating conditions by interpolating within the factor levels actually tested.

Design Flowchart



Design Plan

- An effective design plan should contain the following:
  1. A **clear** description of the objective(s).
  2. An **appropriate** design plan that guarantees unconfounded and unbiased factor effects.
  3. A method for collecting data that will allow for the estimation of the variance of the experimental error.
  4. A stipulation to collect **enough** data to satisfy the objective(s).

Checklist

1. Define Objectives

- Define the objectives of the study.
- Why is the experiment to be performed?
- Classify sources of variability or study cause and effect relationships?
- If it is cause and effect: screening or optimization experiment?
- How large an effect should be in order to be meaningful to detect?

2. Identify Experimental Units

- Identify the item upon which something will be changed.
- Human, animal, raw material for production, or conditions that exist in a point in time?
- Helps with understanding the experimental error and its variance.

3. Define a Response/Dependent Variable

- Define a **meaningful and measurable response or dependent variable.**
- Which characteristics of the experimental units can be measured after each run.
- This characteristic should best represent the expected differences to be caused by changes in the factors.

4. List Independent & Lurking Variables

- Declare which independent variables will be studied.
- Be sure that the independent variables selected can be controlled in a single run and varied from run to run.
- Variables that are thought to affect the response, but cannot be controlled, are classified as lurking variables.
  - The plan should prevent the lurking variables from biasing the study.

5. Run Pilot Tests

- Run pilot tests to be sure that the factors can be controlled, the response can be measured, and the replicate measurements are similar.
- If there is a problem during the pilot tests go back to the previous steps.
- If the pilot tests are successful the study will produce usable data.

6. Flow Diagram of the Experimental Procedure

- **Make a flow diagram of the experimental procedure for each run.**
- Helps to ensure the procedure is understood and standardized across all of the runs.

7. Select the Experimental Design

- Select an experimental design that suits the objective(s) of the experiment.
- Includes:
  - Description of which factor levels will be studied.
  - How the experimental units will be assigned to the factor levels or combination of factor levels.
- The choice of experimental design will determine which model will be selected for analysis.

8. Determine the Number of Replicates Required

- Determine the number of replicate runs that will give a high probability of detecting an important effect.
- This number is based on the expected variance of the experimental error and the size of a practical difference.

9. Randomize the Experimental Conditions to Experimental Units

- Based on the selected design, the method for randomly assigning experimental conditions to the experimental units is outlined.
- The way the randomisation is done determines the way the data should be analysed.
- It is important to describe and record exactly what has been done.

10. Describe a Method for Data Analysis

- Outline the steps for the analysis.
- Analysing simulated data can often help to verify the selected methods will work.

11. Budget & Timetable for Resources Needed

- Create a budget and timetable for the resources needed to complete the experiments.
- Having a schedule may improve the chances of completing the research on time.
- A budget will help to determine the feasibility of the research and allow for any applicable funding request.

- Careful planning and execution of that plan are the most important steps.
- Proper planning will reduce the probability of problems arising later on.
- There are many things to consider when performing the experiments.

Illustrative Example 2

- An experiment was carried out on 24 tomato plants to determine the effects of watering levels and fertilizer on the mass of the tomatoes.
- Two different fertilizers (Fertilizer A, & Fertilizer B) were used.
- Three different water levels were used (100ml, 200ml, 400ml per day).
- Identify the experimental units (subjects), factors (treatments), levels, response variable, and how to design the experiment.

Illustrative Example 2

- Experimental units: Tomato plants
- Factors:
  - Fertilizers
  - Water volumes
- Levels:
  - Fertilizers: A and B
  - Water volumes: 100ml, 200ml, 400ml
- Dependent variable: Mass of tomatoes
- Possible design:
  - 6 total combinations of factors
  - Split the plants into groups of 4
  - Apply each of the possible combinations to one of the groups

Illustrative Example 2 Considerations

- Possible problems with this design?
- Lurking variables?
- Implications of the results?

Design Selection

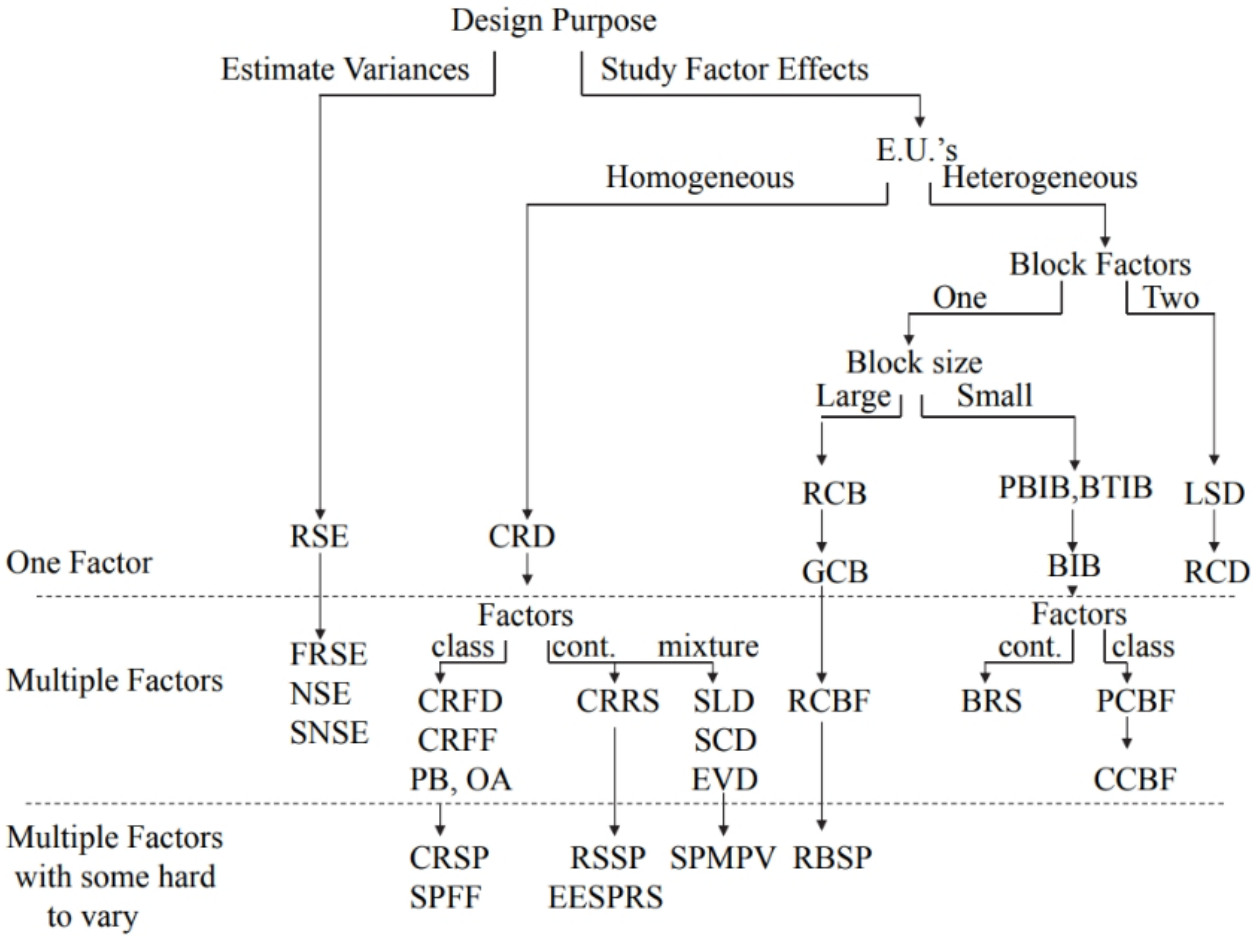


Figure: Source: (1)

Design Selection Index I

Notation	Meaning
RSE	Random sampling experiment
FRSE	Factorial random sampling experiment
NSE	Nested sampling experiment
SNSE	Staggered nested sampling experiment
CRD	Completely randomized design
CRFD	Completely randomized factorial design
CRFF	Completely randomized fractional factorial
PB	Plackett-Burman design
OA	Orthogonal array design
CRSP	Completely randomized split plot
RSSP	Response surface split plot
EESPRS	Equivalent estimation split-plot response surface

Design Selection Index II

Notation	Meaning
SLD	Simplex lattice design
SCD	Simplex centroid design
EVD	Extreme vertices design
SPMPV	Split-plot mixture process variable design
RCB	Randomized complete block
GCB	Generalized complete block
RCBF	Randomized complete block factorial
RBSP	Randomized block split plot
PBIB	Partially balanced incomplete block
BTIB	Balanced treatment incomplete block
BIB	Balance incomplete block
BRS	Blocked response surface
PCBF	Partially confounded blocked factorial
CCBF	Completely confounded blocked factorial
LSD	Latin-square design
RCD	Row-column design

Exercise 1

- Explain the difference between an experimental unit and a sub-sample or sub-unit.
- Explain the difference between a sub-sample and a duplicate.

Exercise 2

- Describe a situation within your realm of experience (your work, your hobby, or school) where you might like to predict the result of some future action.

- Explain how an experimental design, rather than an observational study, might enhance your ability to make this prediction.

References & Resources

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

- The largest and most expensive medical experiment in history
- Glossary of experimental design