Module 1
Fundamentals of Statistics

Module Learning Outcomes

- Understand the composition of and provide a description of an objective.
- Provide a description, in words, the subjects of interest.
- Provide a description, in words, the variable(s) of interest. Include units of measurement for numerical variables or list out all possible classes for categorical variables.
- Identify the type of the given variable(s).
- Identify the scale of measurement of the given variable(s).
- Identify the use or the role of two variables.

What we do in (the discipline of) Statistics?

- Statistics starts with an idea or a question – like I wonder if my hypothesis was right, or I wonder if these two things were related etc.
- We will not need (to do) statistics when the information (and hence answer/result) is readily available to you (like you can find answers on Google).
- In other words, we only do statistics when we want to know something that nobody can tell you or answers are not available on Google. In other words, we will need to figure out the answer ourselves by doing a **study** or **statistical study**.

The "WH" Questions

- Just like writing an essay, it is very important to have an outline or a plan by asking the "wh" questions.
- "Who" refers to the subjects of interest.
- "What" refers to the variables of interest.
- "When" and "where" refer to the location and time/date the research is done or the data is collected.
- "Why" refers to the motivation of this study, or simply the objective or goal.
- "How" is a million-dollar question, as it involves what you need to do to answer the Why. But it refers to how to get the sample (or which sampling method to use) and how the data is collected in this course.

1.1 Objective

- In this course, we will learn how to methodically convert questions (or speculations) into answers (or results), by going through a few key steps.
- The first step is to convert the idea or question into a form of a workable statement and it is called the **objective** or the **goal** of the study.
- And objectives should contain something can be measured readily.

Examples of Ideas	Examples of Workable Statements
• *Suppose we would like to compare the English competency of foreign students by their county or origin in the Fall semester.*	• *We want to find out how the TOEFL scores vary by the country of origin among all foreign students in the Fall semester.*
• *Suppose we want to compile the ranking of figure skaters in the current competitive season.* *Mind you that they may not compete against each other in all competitions.*	• *We want to compare the average IJS (International Judging System) scores of figure skaters in all competitions of the current season.*

Note that the underlined parts are something that could be measured.

- A proper objective should contain the Who, What, When, and Where.
- MLO: Understand the composition of and provide a description of an objective.

1.2 Subjects (of a Study)

- **Subjects** refer to the "things" that we are interested in.
- Other common synonyms are "**individuals**", "**cases**", "**items**", "**elements**" etc.
- Subjects often refer to solid things like people, houses, and cars etc. But they could also be intangible things like days.
- If they are available, description of subjects often includes the Where and When.
- Good-to-know: Subjects are always in plural form (i.e. always has –s at the end) because we rarely are interested in one person or one car.
- Note: Your textbook uses a different word, but I will always use subjects in the notes.
- MLO: Provide a description, in words, the subjects of interest.

1.3 Variables (of a Study)

- **Variables** are the direct characteristics of the subjects.
  Example: Number of bedrooms is a direct characteristic of houses. Therefore, the Number of Bedrooms is the proper variable of the subject Houses.
  Example: Consider the memory capacity of a smart phone owned by a Langara student. Memory capacity is not a direct characteristic of Langara students. Therefore, Memory Capacity cannot be the variable of the subjects Langara Students.
- **Variable Name** is the name of the variable, typically no more than three words. (Just like our names do not usually have more than three syllables.) It allows us to make a quick reference to them.
  Example: "Number of Bedrooms" and "Memory Capacity" are proper variable names, but "The Number of Bedrooms of Houses in Vancouver City" and "The Amount of Memory a Smart Device has" are not good variable names.
- **Variable Description** is a detailed version of the variable name.
  Sometimes, the variable name itself is self-explanatory. So, the name and description could be the same. Otherwise, a longer and more detailed description of the variable is required.
  Example: The variable name "Opinion" itself does not mean much without the description like "the opinion (of Langara students) about the quality of instructions". In contrast, the variable name "Price" (of houses) is self-explanatory and does not need any further description.
- The description of categorical variables is a bit tricky sometimes, especially when there are only two responses (like Yes vs. No). In this case, we would like to use the combo "**whether or not**…"
- An observed answer of variable is called a **response**.
- MLO: Provide a description, in words, the variable(s) of interest. Include units of measurement for numerical variables or list out all possible classes for categorical variables.

1.4 Data set

- A data set is an organized form of two key components – subjects and variables.
- Each subject (of the study) occupies a row; each variable takes a column.
- A response is simply an intersection between a row and a column in the data set.
- The **sample size** (or the number of subjects in the sample) determines the number of rows and the number of variables determines the number of columns in a data set.
- The term "data set" (note there are two words) is the proper one, but others like "dataset" (one-word version) and simply "data" are universally accepted.
  Example
  Note: Only the grey-out part is the data set. The rest is only my annotation.

		Variables				
		Var #1	Var #2	Var #3	Var #4	Var #5
	Name of Variable →	Price	Age	Location	Bedroom	View
Subjects	Subject #1	1240000	28	Kitsilano	4	Yes
	Subject #2	1580000	55	Oakridge	4	No
	Subject #3	950000	75	Marpole	6	No
	Subject #4	920000	25	Killarney	3	No
	⋮	⋮	⋮	⋮	⋮	⋮
	Subject #100	1350000	43	Oakridge	4	No
	Sample Size (n = 100)					
	Unit of Measurement	$	(years old)	None	(number of)	None

- Note: Units are never entered in the data set (e.g. no dollar sign "$" under Price) because statistical software applications typically have trouble handling numbers and units together. Therefore, it is your job to keep track of the units in other places like the description of variables in the documentation.

1.5 Type of Variables

- There are two **types** of variables – categorical variables vs. numerical variables.
- There are two questions to ask yourself before determining the type of a variable.
- Question #1: Are the responses in numbers or non-numbers?
  If the responses are non-numbers, it is definitely a **categorical variable** (or **qualitative variable**). Otherwise, it is typically classified as **numerical variable** (or **quantitative variable**), but make sure you check question #2 before confirming.
- Question #2: Are there strictly more than 10 distinct numbers in the responses?
  If so, then it can be considered a numerical variable.
  Otherwise, it has to be considered as a categorical variable (or it is sometimes called discrete variable).

- Typically, numerical variables have **unit of measurements**, or simply **units**, like unit *kilograms* goes with the numerical variable *Weight*; or *metres* goes with *Height* etc. A very good indicator is to check if the variable has units. If so, it is very likely to be a numerical variable.
- In a proper study, we should know (or be able to list out) all possible responses of any categorical variable. The distinct responses are called the **classes** or **categories**.
- Both the unit of measurements (for numerical variables) and the list of classes (for categorical variables) are essential component when describing variables.
- MLO: Identify the type of the given variable(s).

1.6 Scale of Measurement

- The **scale of measurement** or simply **scale** (do not confuse this with units of measurement) is not a terribly important thing to know in life.
- This is basically a sub-division of the type of variable.
- There are two options for numerical variables: **interval scale of measurement** or **ratio scale of measurements**. Because of the subtle difference, we are not trying to differentiate the two in the course. In other words, we group them as **interval/ratio scale of measurement**.
- There are also two options for categorical variables.
- 1) **Ordinal scale of measurement** or **ordinal scale**: Responses in this class can naturally be ranked or ordered in some sensible way, without referring to other attributes.
  Example: The five options of Satisfaction Rating (with options Extremely Satisfied, Moderately Satisfied, Neutral, Moderately Dissatisfied and Extremely Dissatisfied) are in natural order.
- 2) **Nominal scale of measurement** or **nominal data**: Responses in this class cannot be naturally ranked or ordered in any sensible way.
  Example: Eye colour, religion, ethnicity etc.
- MLO: Identify the scale of measurement of the given variable(s).

1.7 Use of Variables

- While the type and scale of variables are needed for each variable in every study, the **use of variables** are needed in studies with two or more variables (which means "always").
- If the word "use" confuses you, you could think of it as the role of the variable plays in a study.
- When there are two variables in a study, we will need to tell them apart: one is called the **independent variable** and the other is called the **dependent variable**.
- There are other definitions and you will get better at it with experience, but here is a good start:
- A variable (or multiple variables) that will be used as the role of **independent variable** when it is about demographics or background information (like gender, birthplace, height, number of siblings etc.), or something that does not change often.

- A variable that will be used as the role of **dependent variables** when their responses are mostly about opinion of subjects (like what you think about the government's recent scandal etc.) or something that changes more frequently.
- Like I said, the variety of definition is endless. This is only a basic guideline.
- Note: Keep in mind that it all depends on which variable it is paired up to and the role could be totally different.

Example Objective: To find out how GPA is dependent on the amount of gaming time among students. Dependent Var: GPA *Independent Var: Amount of gaming time*	Example Objective: To find out how different genders of students have different amount of gaming time. *Dependent Var: Amount of gaming time* Independent Var: Gender

- Another way of calling them are the **response variables** and the **explanatory variables**.
- MLO: Identify the use or the role of two variables.