Basic Terminology

Reading

Sections 1.1, 1.2

Practice Problems

1.9.1 (Page 55) 1.1, 1.2
1.9.2 (Page 56) 1.4, 1.5, 1.6, 1.8

Notes

- The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**.
- **Descriptive Statistics** deals with organizing and summarizing of the data
- **Inferential Statistics** deals with drawing conclusions from sample data about the larger population they represent.
- Key Terms:

  Population A collection of "things/entities/persons" we want to study. Often we are unable to collect information for all members of the population. We rely on a representative sample.

  Sample A *manageable* portion/subset of the population. We collect information only for the sample, and this gives us our data. How the sample is to be selected is an important question to consider.

  Statistic A statistic is a number that represents a property of the sample data. For instance if we have a sample of students, it could be their average gpa.

  Parameter A number that is a property of the entire population. Typically unknown to us. But we use the statistic to try to estimate it.

  Variable A characteristic of each individual/case in the population. For each individual and each variable there is a corresponding well-defined and unique value.

  Data The actual values of the variable, one for each individual.

- **Variables** are arranged in two main types:

  Scalar Also called **numerical** or **quantitative** variables. Their values are numbers in a scale, in some specific unit of measurement. A defining characteristic of scalar variables is that it makes sense to form averages. Examples: GPA, height in feet, income in USD.

  Scalar variables are often broken into two types: **continuous** and **discrete**. For continuous variables all real numbers in some range are possible values.

For discrete variables, only certain numbers are allowed. For instance the number of family dependents, or a county's population, can be thought of as discrete. A person's height would be continuous.

Categorical Also called **qualitative**. They classify the individuals in groups. Examples: Gender, Grade.

Categorical variables are further divided into **Nominal** and **Ordinal**, depending on whether the different categories have a natural order to them or not. Gender would thus be nominal, Grade would be ordinal.

- Some times it may be hard to determine if some data is discrete or ordinal. The main difference is the types of analyses we do and questions we ask. In general, if the distance between the numbers is important, that tends to make the values scalar. If on the other hand there are too few different possible values, and their relative ordering is the only important factor, that is a reason to treat the variable as ordinal.
- Activity: List a number of different variables that you could measure on students. Then assign a type to each variable.
- **Relationships between variables**. Usually multiple variables are measured on the same individuals. We can then ask how these variables relate to each other. This will be an important component of the class.

Dependent Two variables are called **dependent**, or **associated**, or **related**, if they show some connection to each other; in other words if knowing the value of the one variable on an individual can give us some information about the value of the ohter variable on that individual.

An example for instance would be average GPA and gender. In general female students tend to have a higher GPA than male students. So knowing a specific student's gender gives us some information about their GPA, but it will not completely determine their GPA.

When variables are quantitative and associated, then we can talk about a **positive** or **negative association**. In a positive association the large values of the one variable tend to be paired up with the large values of the other variable, and the small values of the one variable tend to be paired up with the small values of the other variable. For a negative association it is the other way around.

Independent Variables that are not related are called **independent**. In that case knowing the value for one variable on a specific individual does not convey any information about the value of the other variable.