Statistics for the Biological, Environmental and Health Sciences

STAT 007

Introduction to Statistics

Chapter 1

Types of Data

Section 1-2

- In this section we will:
  - define the terms *statistic* and *parameter*.
  - discuss different types of data.

- A major use of statistics is to collect and to use *sample data* to make conclusions about the *population*.
- Importance of knowing type of data: Different types of data require different statistical analysis.

Type of Data
categorical v/s quantitative
discrete v/s continuous
nominal, ordinal, interval or ratio level of measurement

- Important: Don't do computations and don't use statistical methods that are not appropriate for the data.

Parameter and Statistics

Definition

*A **parameter** is a numerical measurement describing some characteristic of the population.*

*A **statistic** is a numerical measurement describing some characteristic of a sample.*

Example

Identify a *parameter* and a *statistic* in the following statement:

"There are 17,246,372 high school students in the United States. In a study of 8505 U.S. high school students 16 years of age or older, 44.5% of them said that they texted while driving at least once during the previous 30 days."

Types of Data: Categorical v/s Quantitative

Definition

- ***Categorical*** *(or qualitative or attribute)* ***data*** *consist of names or labels or categories (not numbers that represent counts or measurements).*
- ***Quantitative*** *(or numerical)* ***data*** *consist of numbers representing counts or measurements. It should include appropriate unit of measurement.*

Types of Data: Categorical v/s Quantitative

Example

Discuss whether type of data reported below are categorical or quantitative:

a) The gender (male / female) of subjects enrolled in a clinical trial.

b) The identification numbers 1, 2, 3,..., 25 assigned randomly to 25 subjects in a clinical trial.

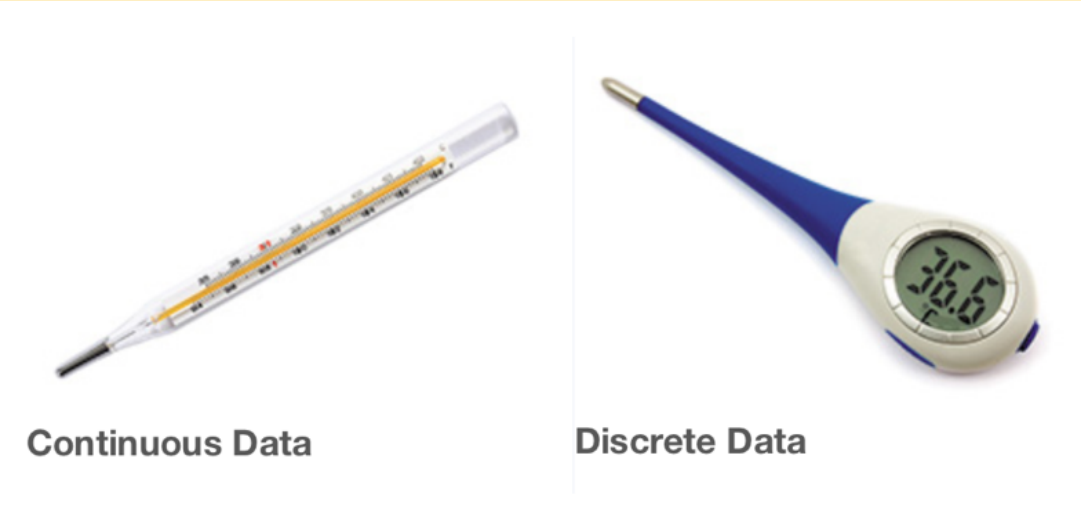
c) The age (in years) of the subject enrolled in a study.

d) The systolic blood pressure (in millimeters of mercury (mmHg)) of the subject enrolled in a study.

Types of Data: Discrete v/s Continuous

Definition

- ***Discrete data***: *data values are quantitative and number of values is finite or "countable" (there are infinitely many values).*
- ***Continuous data***: *infinitely many possible quantitative data values. The collection of values is not "countable" (values are on a continuous scale).*



Types of Data: Discrete v/s Continuous

Example

Discuss whether type of data reported below are discrete or continuous.

a) Number of physical examinations a subject enrolled in a study, that lasts one year, has.

b) The volume of blood drawn from a subject (between 0 mL and 50 mL).

c) Suppose that the accuracy of a blood typing test wants to be tested. To do this, a sample of the subject's blood type is submitted to the test until the test yields an error. How many times it is submitted is recorded.

Types of Data: Level of Measurement

Definition

- *The **nominal level of measurement data** type is not possible to arrange in some order.*
- *The **ordinal level of measurement data** type can be arranged in some order, but differences (obtained by subtraction) between data values either cannot be determined or are meaningless.*
- ***Interval level of measurement data** can be arranged in order, differences between data values can be found and are meaningful; but data at this level do not have a natural zero starting point at which none of the quantity is present.*
- ***Ratio level of measurement data** can be arranged in order, differences can be found and are meaningful, and there is a natural zero starting point (where zero indicates that none of the quantity is present). Also the term "twice" makes sense.*

Types of Data: Level of Measurement

Example

Discuss whether the type of data are of the nominal, ordinal, interval or ratio level of measurement.

a) Survey responses of “yes”, “no”, and “undecided”.

b) Letter grades of A, B, C, D, or F in a biostatistics class.

c) The body temperature (in degrees Farenheit) of the subject.

d) Year when the subject enrolled in the study.

e) The height (in cm) of a plant.

Practice

Look at the exercises at the end of Section 1-2 in page 22.

Specially, look at exercises:
1, 2, 3, 4, 5, 7, 9, 11, 12, 13, 15, 17, 19, 21, 23, 25, 26, 27, 28, 30, 31, 33.

Collecting Sample Data

Section 1-3

- In this section we will:
  - Describe methods for collecting data sets.
  - Describe methods for collecting samples.
  - Describe different types of data collection methods.

- Recall that **Statistics** is the science of planning studies and experiments; obtaining data; and organizing, summarizing, presenting, analyzing, and interpreting those data and then drawing conclusions based on them.
- In a statistical study, it is very important to use an appropriate method for collecting the sampling data. The two methods that we will discuss are *experiment* and *observational study*.
- We will introduce several *sampling methods*. Particularly important is the sampling method of using a *simple random sample*.

"If sample data are not collected in an appropriate way, the data may be so utterly useless that no amount of statistical torturing can salvage them."

Data Collecting Methods

Definition

a) ***Experiment***: *In an experiment, we apply some treatment and then proceed to observe its effects on the experimental unit.*
   - ***Experimental Units***: *the individuals in an experiment. When the individuals are people, the experimental units are called* ***subjects***.
   - *Can claim causation.*
   - *Good Designs of Experiment include replications, blinding (or double blinding), and randomization.*
   - *It is not always possible to perform an experiment.*

b) ***Observational Study***: *In an observational study, we observe and measure specific characteristics, but we don't attempt to modify the individuals being studied.*
   - ***Lurking Variable***: *is a variable that affects the variables included in the study, but is not included in the study.*
   - *Can only claim association, not causation.*

Common Sampling Methods

Definition

- *A **simple random sample** of n subjects is selected in such a way that every possible sample of the same size n has the same chance to be chosen.*
- *In **systematic sampling**, we select some starting point and then select every kth (such as every 4th) element in the population.*
- *In **convenience sampling**, we simply use data that are very easy to get.*
- *In **stratified sampling**, we subdivide the population into at least two different subgroups (or strata) so that subjects within the same subgroup share the same characteristics (such as gender). Then we draw a random sample from each subgroup (or stratum).*
- *In **cluster sampling**, we first divide the population area into clusters (or sections). Then we randomly select some of those clusters and choose all the members from those selected clusters.*

Types of Observational Studies

Definition

- *In a **cross-sectional study**, data are observed, measured, and collected at one point in time.*

  Comment: cheap and easy to do, but they don't give very strong results.

- *In a **retrospective (or case-control) study**, data are collected from a past time period by going back in time (through examination of records, interviews, and so on). Typically, one group of the subjects has a characteristic (case group) and the other group does not have the characteristic (control group).*

  Comment: individuals need to have an accurate memory and also respond honestly.

- *In a **prospective (or longitudinal or cohort) study**, data are collected over a period of time in the future from groups that share common factors (called cohorts). Usually, they aim to determine if particular characteristics affect a variable of interest.*

  Comment: most powerful of the observational studies.

Types of Experimental Designs

Definition

- ***Completely Randomized Experimental Design***: *Assign subjects to different treatment groups through a process of random selection.*
- ***Randomized Block Design***: *First, blocks (or groups) of subjects with similar characteristics are formed. Second, treatments to the subjects within each block are randomly assigned.*

  Comment: avoids possible sources of ***confounding****.*
- ***Matched Pair Design***: *Compare two treatment groups by using subjects matched in pairs. The following are examples of matched pairs:*
  - *Twins: people that is very similar and each receives a different treatment.*
  - *Before/After: the same subject is measured before and after some treatment.*

Sampling Errors

Definition

- *A **sampling error** (or random sampling error) occurs when the sample has been selected with a random method, but there is a discrepancy between a sample result and the true population result.*

  Comment: such an error results from chance sample fluctuations.

- *A **nonsampling error** is the result of human error, such as wrong data entries, computing errors, questions with biased wording, false data provided by respondents, forming biased conclusions, or applying statistical methods that are not appropriate for the circumstances.*

- *A **nonrandom sampling error** is the result of using a sampling method that is not random, such as using a convenience sample or a voluntary response sample.*

Practice

Look at the exercises at the end of Section 1-3 in page 31.

Specially, look at exercises:
1 to 36.