Fundamentals of Statistics

Dr. Wandi Ding

Chapter 1 data collection

1.1 Introduction to the statistics

- what is statistics?

The statistics is the science of data, but in reality statistics is both the science and the art of collecting, organizing, summarizing, analyzing and drawing conclusion based on data. (different books have different definitions but similar).

CHAPTER 1 DATA COLLECTION

1.1 INTRODUCTION TO THE STATISTICS

- what is data?

Data is the values (measurements or observations) that the variables can assume. A collection of data values forms a data set, each value in the data set called a data value or a datum, or one observation,individual.

CHAPTER 1 DATA COLLECTION

1.1 INTRODUCTION TO THE STATISTICS

- What is variable?

A variable is a characteristic or attribute that can *__assume different values.__* *(If a variable did not vary and has the same values , it will be a constant.)*

EXAMPLE: daily temperature , each person's weight in our classroom, the model of cars on campus of MTSU…

Your current weight is a variable, right?

False! Why?

CHAPTER 1 DATA COLLECTION

1.1 INTRODUCTION TO THE STATISTICS

Classification of variables:

- Independent vs. dependent variables (classification 1)

  1. independent variable in an experimental study is the one that is being manipulated by the researcher, also called explanatory variable.
  2. dependent variable is the resultant variable, also called outcome variable (response variable).

Ex. To see whether there are cause-effect relationship between the grade (dependent) and attendance (independent).

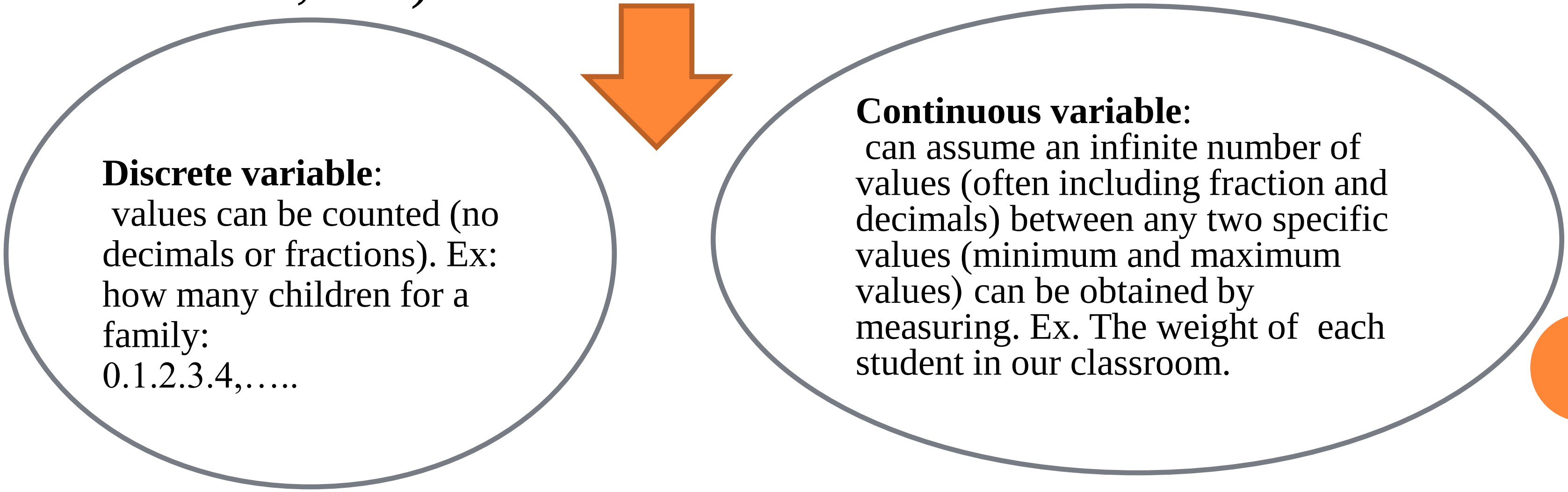
Chapter 1 data collection

1.1 Introduction to the statistics

- Qualitative vs. quantitative variable (classification 2)

1. qualitative variables involve attributes (such as gender, occupation, or other category) also called categorical variables.

2. quantitative variables involve numerical value, could be ordered or ranked. (such as age, weight, height, income, etc.).



Discrete variable:
values can be counted (no decimals or fractions). Ex: how many children for a family:
0.1.2.3.4,…..

Continuous variable:
can assume an infinite number of values (often including fraction and decimals) between any two specific values (minimum and maximum values) can be obtained by measuring. Ex. The weight of each student in our classroom.

CHAPTER 1 DATA COLLECTION

1.1 INTRODUCTION TO THE STATISTICS

- A data set of a small company with 6 workers:

? ? ? ?

Name	Annual income	# of children	Gender	Married status
Mike	51.5k	1	M	Y
Jessica	41k	2	F	Y
Kevin	61.2k	0	M	N
Tobey	20k	1	M	Y
Meredith	55k	0	F	N
Tom	37.5k	2	M	Y

Qualitative data

Q: Qualitative variable? Quantitative variable? Discrete variable? Continuous variable?

A: qualitative (name, gender and married status),quantitative variable (annual income—continuous, # of children-discrete).

CHAPTER 1 DATA COLLECTION

1.1 INTRODUCTION TO THE STATISTICS

- Level of measurement on variables (classification 3)

1. Nominal data (nominal level of measurement) classifies data value into mutually exclusively categories (non-overlapping) in which no order or ranking.

Ex. Gender, name , etc.

Nominal data can be imposed on the data. In other words, they are numbers representing arbitrary code.

Ex. If you use number 1 to indicate male, and 2 indicate female. Thus, 1,2 are nominal data.

Can you give me another example of nominal data?

1.1 INTRODUCTION TO THE STATISTICS

- Level of measurement on variables (classification 3)

2. Ordinal data (ordinal level of measurement) classifies data value into categories that can be ranked; however, precise difference between the ranks do not exit. In other words, ordinal data convey ranking in terms of importance, strength performance , severity, etc.

Ex. A value of 3 indicate a gentle sea breeze; 6 represents a strong breeze; 9 signifies a strong gale; however, the change in force between a strong sea and a gentle breeze IS NOT EQUAL TO that between a strong gale and a strong breeze.

Another example?

A B C D letter grade earned on this class.

CHAPTER 1 DATA COLLECTION

1.1 INTRODUCTION TO THE STATISTICS

- Level of measurement on variables (classification 3)

3. Interval data (interval level of measurement) can be ranked, the precise differences between units exist; and ratios between units do not have meaningful. A value of zero does not mean the absence of the quantity.

Ex. Temperature, people’s IQ

Note: only the addition and subtraction of the arithmetic operations are valid for interval data.

1.1 INTRODUCTION TO THE STATISTICS

- Level of measurement on variables (classification 3)

4. Ratio data (ratio level of measurement) possesses all the characteristics of interval data, and true ratios exist and have meaning. A value of zero means the absence of the quantity.

Ex. Weight, height, salary, etc.

Note: the addition, subtraction, multiplication and division are all valid for ratio data.

1.1 INTRODUCTION TO THE STATISTICS

- Two major fields of statistics:

1. Descriptive statistics:

consists of collecting, classifying, organizing, summarizing and presentation of data.

2. Inferential statistics:

consists of generalizing from samples to populations, performing estimations and hypothesis test, determining relationships among variables, making predictions and measure the reliability of the result.

1.1 INTRODUCTION TO THE STATISTICS

- Note about inferential statistics:

(1) inferential statistics uses probability as a basis.

(2) Making reference on sample about population.

- **Sample**---is a group of subjects selected from a population.
- **Population**---consists of all subjects which are being studied.
- A **parameter** is a summary number about some characteristic of the population.
- A **statistics** is a summary number about some characteristic of the sample.

Chapter 1 Data Collection

1.1 Introduction to the Statistics

- **Ex.** Suppose we want to study the proportion of students on campus of MTSU with a cell phone; and we get a value, 98%. But for our classroom, the proportion is 100%. Which value is a parameter? Which is a statistics? Why?
- **Validity** of a variable or measurement indicates how close to the true value the measurement is.
- **Reliability** of a variable or measurement represents the ability of different measurements of the same individual to yield the same results.

1.2 OBSERVATIONAL AND EXPERIMENTAL STUDY

- **Observational study**

Measuring the value of the response variable without influence the value of either the response or explanatory variable.

The researcher merely observes what is happening or what has happened in the past and draw the conclusion from the observations.

- **Experimental study**

In a study if a researcher assigns the individuals to a certain group, intentionally changes the values of an explanatory variable and then record the value of the response variable for each group.

The researcher manipulates one of the variables and to see how the manipulation variable influences other variables.

CHAPTER 1 DATA COLLECTION

1.2 OBSERVATIONAL AND EXPERIMENTAL STUDY

- **Confounding variable (lurking variable)**

  is an explanatory variable that was not considered in a study but that affects the value of the response variable in the study. Typically, confounding variables are related to explanatory variables in this study.

  For example, A doctor want to check if Drug A has an effect on causing headache. The doctor must design an experiment to make sure that the subjects in the study felt headache only because of the influence of the Drug A, and not caused by other factors. Those other factors would be confounding variables.

1.2 OBSERVATIONAL AND EXPERIMENTAL STUDY

- Three major observational studies:

1. Cross-sectional study:

observations are collected at a specific point in time or over a very short period of time.

2. Case-control study:

this kind of study is retrospective, and require individuals to look back in time and look at the existing records. In this study, individuals that have a certain characteristic are matched with those that do not.

3. Cohort study:

this kind of study is prospective, first identify a group of individuals to participate the study (the cohort). The cohort is then observed over a period of time. During this period, characteristics about the individual are recorded and some individuals will be exposed to certain factors (not intentionally) and others will not.

1.2 OBSERVATIONAL AND EXPERIMENTAL STUDY

- Existing sources of data: (data available to public)

1. www.cdc.gov (centers for disease control and prevention)
2. www.irs.gov (the internal revenue sevice)
3. http://fjsrc.urban.org/index.cfm (the department of justice)
4. www.gss.norc.org (general social survey)
5. Another source of data is census

A **census** is a list of all individuals in a population along with certain characteristics of each individual.

1.3 SIMPLE RANDOM SAMPLING

- **Random sampling:**

  is the process of using chance to select individuals from a population to be included in the sample.

- **Suppose**: the local Kroger store manager wants to check the extent of the customers to the store. A sample of the first 200 customers will be the representative of the population? The sample is random sample? Why?

Answer: this sample is not the representative of the population and not a random sample because the individuals in this sample is not selected using chance.

CHAPTER 1 DATA COLLECTION

1.3 SIMPLE RANDOM SAMPLING

- There are four basic sampling techniques: simple random sampling, stratified sampling, systematic sampling and cluster sampling.
- **Simple random sampling:**

  A sample of size n from a population of size N is obtained through simple random sampling if every possible sample size n has an equally likely chance of occurring. The sample is then called a simple random sample.
- **Sample with replacement:** which means the selected individual is place back into the population. With the method, some individuals could be chosen more than one time, some maybe could be not chosen even one time.
- **Sample without replacement:** which means once the individual is chosen, and the individual will not place back into the population.

1.3 SIMPLE RANDOM SAMPLING

- Ex. How to get a simple random sample?

1. Let the letters of A, B, C indicate a population and we want to choose a random sample of size 2.

2. List all possible sample of size =2; (A,B) (A,C) (B, C), totally there are 3 random samples with size=2.

3. For the 3 random samples with size=2, each random sample has the same chance to be chosen.

4. Which sample will be chosen? Random number will help us. There are some ways to generate random numbers to help choose sample.

CHAPTER 1 DATA COLLECTION

1.3 SIMPLE RANDOM SAMPLING

- Ex. Let’s go over the class activity on page 23 and example 2 on page 24.

- From these two examples, Know how to use random number table and understand the idea of simple random sampling.

- All statistical software can have random number generator function, such as SAS, NCSS, Minitab, SPSS,JMP, etc.

CHAPTER 1 DATA COLLECTION

1.4 OTHER SAMPLING METHODS

- Stratified sampling: is obtained by separating the population into nonover-lapping groups called strata and then obtaining a simple random from each stratum. The individuals within each stratum should be homogeneous (similar) in some way.
- Systematic sampling: is obtained by selecting every kth individual from the population. The first individual selected corresponds to a random number between 1 and k.
- Cluster sampling: is obtained by selecting all individuals within a randomly selected cluster.
- Figure 5 on page 35 shows a good summary on these sampling techniques.

CHAPTER 1 DATA COLLECTION

1.5 BIAS IN SAMPLING

- Bias: if the results of sample are not representative of the population, then the sample has bias.
- Three sources of bias in sampling:

1. sampling bias: means that the sample tends to favor one part of the population over another.
2. non-response bias: when individuals selected to be in the sample who do not respond to the survey have different opinions from those who do.
3. response bias: when the answer do not reflect the true feelings of the respondent.

Chapter 1 Data Collection

1.5 Bias in Sampling

Non-sampling error:

That results from under Coverage, non-response bias, response bias, or data-entry error. Such errors could also be present in a complete census of the population.

Sampling error:

That results from using a sample to estimate information about a population. This type of error occurs because a sample gives incomplete information about a population.