Packet 2: Introductory Definitions

Textbook pages: 1 – 8; 267 – 277 (selected topics)

After completing this material, you should be able to:

- discuss what statistics is and how it can be applied to specific examples.
- identify the **population, sample,** and what a **statistical inference** is for a specific example.
- identify **statistics** that are calculated for a specific example.
- identify the variable recorded for a specific example and classify the variable as categorical or quantitative.
- identify the sample selected and determine if the sample is a simple random sample, a convenience sample, or a voluntary response sample

How can we define **statistics**?

2 broad areas of statistics:

1.

2.

What exactly is **data** and why is it so important?

Data is collected when characteristics are recorded about each individual – these are referred to as **variables**. There are **two types of variables**:

1.

2.

Knowing what type of variable is measured will prove to be *very* important – the type of variable determines how the variable can be summarized, as well as what type of inference can be conducted.

In the space below, list several examples of categorical and quantitative variables:

Example 1: A food retailer that specializes in selling organic food has decided to open a new store. To help determine the best location in the US for the new store, researchers decide to examine data from current stores, including weekly sales ($), town population (thousands), median age of the town, median income of the town ($), and whether or not the store sells wine and beer. For this scenario, identify as many of the following as possible:

Who:

What:

Why:

Additional Definitions

Population:

Sample:

Parameter:

Statistic:

Example 2: On July 15, 2004, the Harris Poll released the results of a survey asking whether people favored or opposed abolishing the penny. Of a national sample of 2136 adults, 59% opposed abolishing the penny. Each of the following statements using one of the definitions we have discussed incorrectly. Replace the underlined portion with the term or phrase which will make the sentence true.

The **population** is the 2136 adults contacted by the Harris Poll.

The **sample** is the 59% who oppose abolishing the penny.

The variable is **the 59% who oppose abolishing the penny**.

The **parameter** consists of all American adults

It is going to be important to select *samples* from *populations* in order to make statistical inferences. Because of this, we need to understand different ways samples are selected (some of which are good sampling schemes and some of which are poor sampling schemes).

First, we need to define the **sampling frame:**

Types of Samples

Example 3: State legislators are interested in estimating the percentage of Kentucky adults who have a college degree (associate's degree or higher). To investigate, a simple random sample of 200 residents is selected using voter registration records. Each person is contacted and asked whether or not they have a college degree (associate's or higher). After completing the survey, it was found that 43 individuals had obtained an associate's degree or higher.

Who:

What:

Why:

What type of sample was used to select the 200 KY residents? What sampling frame was used?

Identify the population, parameter, sample, and statistic for this example.

Packet 2: Summarizing Categorical Data Textbook pages: 15 – 18

After completing this material, you should be able to:

- summarize categorical data numerically by creating a frequency distribution.
- summarize categorical data graphically by creating a bar graph
- use StatCrunch to create frequency distributions and bar graphs.

Once we have collected data, it doesn't do us much good as simply a dataset – we need ways to organize and display the data in order to glean useful information from it and later to make inferences from the data. Let's start by discussing how to organize and graphical display categorical data.

Organizing Categorical Data

Before we can graphical display a categorical variable, we need to have it organized first. Consider the following example:

The New York City Police Department implemented a "stop-and-frisk" program where officers can stop and question pedestrians and frisk them for weapons and other contraband. In January 2012, 69,073 pedestrians were stopped under the Stop-and-Frisk policy.

The StatCrunch dataset Stop and Frisk, January 2012 contains the following variables – you must first classify these variables are either categorical or quantitative by filling in the table below:

Variable	Categorical or Quantitative?
Sex (coded as 0 = female, 1 = male)	
Race (coded as 1 = black, 2 = black Hispanic, 3 = white Hispanic, 4 = white 5 = Asian/Pacific Islander, 6 = American Indian)	
Age	
Height	
Weight	
FriskOrSearch (coded as 0 = stop did not result in either frisk or search, 1 = the stop did result in a frisk or search)	
FoundSomething (coded as 0 = detainee was not found to have either contraband or weapons, 1 = the detainee had either contraband or weapons)	
ArrestMade (coded as 0 = no arrest was made, 1 = arrest was made)	

The categorical variables can be organized into a **frequency distribution** – when creating a frequency distribution, we essentially "make piles" of common categories in order to see how the cases distribute across categories. When creating a frequency distribution, two common summaries are reported:

frequency:

relative frequency:

A partially filled in frequency table for the frisk or search variable is given below. Using the definitions provided, fill in the missing values.

Frisk or Search?	Frequency	Relative Frequency
None conducted	30,176	
Conducted		

We can get even more specific with the data that was collected. Suppose we restrict our analysis to only those stops where the pedestrian was frisked for a weapon. If a weapon was found, what is the frequency distribution for arrests made? If a weapon was not found, what is the frequency distribution for arrests made?

For detainees where FoundSomething = 1 (yes)

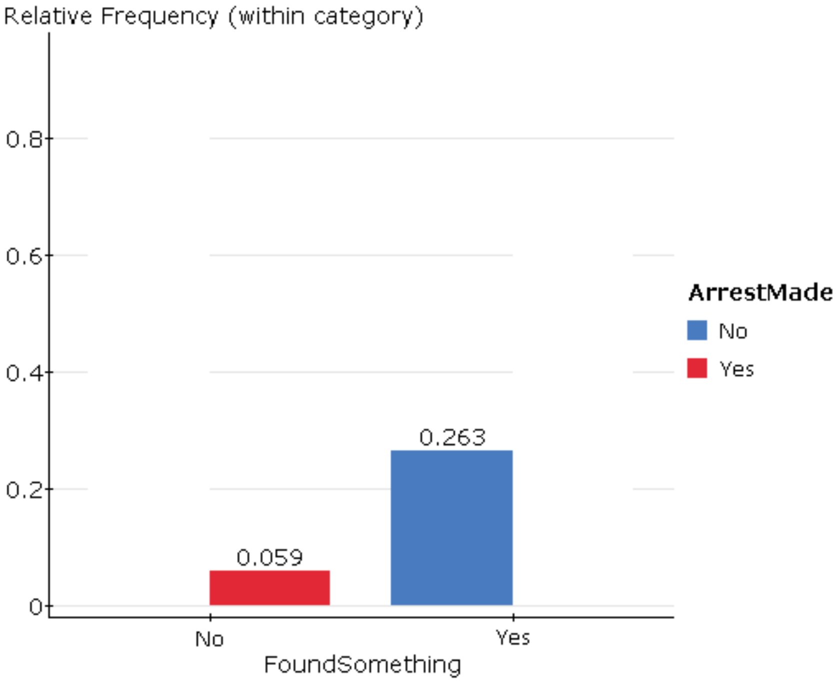
Arrest Made?	Frequency	Relative Frequency
Yes		
No		

For detainees where FoundSomething = 0 (no)

Arrest Made?	Frequency	Relative Frequency
Yes		
No		

While a frequency distribution gives a numerical summary of a categorical variable, it's also nice to have a visual summary. This can be accomplished by creating a **bar graph** – a bar graph displays the distribution of a categorical variable showing either the frequencies or relative frequencies for each category. This allows for a quick visual comparison of categories.

A bar graph comparing arrests made has been partially constructed below. Use the values in the frequency distributions above to finish filling in the graph:


Relative Frequency (within category)
0.8
0.6
0.4
0.2
0
0.059
0.263
No
Yes
FoundSomething
ArrestMade
No
Yes