

Lecture 2

Data Summary/Visualization I

Text: Chapter 2 & Chapter 3

STAT 8010 Statistical Methods I

January 14, 2020

Whitney Huang
Clemson University



Notes

Agenda

- 1 Sampling Techniques
- 2 Summarizing Categorical Data
- 3 Summarizing Numerical Data



Notes

Last Lecture

- Stating the problem, identifying the variable(s) of interest, and gathering data
 - Types of variables and datasets
 - Observational vs. Experimental Studies
 - Methods of sampling
- Summarizing the data
- Analyzing the data
- Reporting and interpreting the results



Notes

Today's Lecture

- Stating the problem, identifying the variable(s) of interest, and gathering data
 - Types of variables and datasets
 - Observational vs. Experimental Studies
 - Sampling Techniques
- Summarizing the data
- Analyzing the data
- Reporting and interpreting the results

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

24

Notes

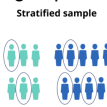
Collecting Data: Statistical Sampling

Statistical sampling is the procedure to select a subset from a statistical **population** that is representative of the population. There are several types of sampling:

- Simple random sampling (SRS): a sample selected such that each element in the population has the same probability of being selected



- Stratified sample: elements in the population are first divided into groups and a simple random sample is then taken from each group



Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

25

Notes

Sampling cont'd

- Cluster sampling: the elements in the population are first divided into separate groups called clusters and then a simple random sample of the clusters is taken that all elements in a selected cluster are part of a sample



- Systematic sampling: randomly select one of the first k elements from the population and then every k_{th} element thereafter is picked



- Convenience sampling: elements selected from the population on the basis of convenience

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

26

Notes

What type of sampling was used?

- ➊ A researcher randomly chooses houses in a town. Once a particular house is chosen everyone living in the house is surveyed
- ➋ A school principal decides to performs an exit interview with every 14th name from a list of graduating seniors
- ➌ A biologist knows that 40% of bats are male and that 60% are female so she randomly selects 20 males and randomly selects 30 females to be in her sample
- ➍ A graduate student wants to do a study on why people like bluegrass music and uses the people she meets at the next show she attends as her sample
- ➎ To get an idea of the average weight of his cattle, a rancher randomly chooses to weigh 25 from his list of the animals

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.7

Notes

Summarizing Categorical Variables

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.8

Notes

Example: Sport Injuries

The paper “*Profile of sport/leisure injuries treated at emergency rooms of urban hospitals.*” by Pelletier et al. 1991 examined the nature and number of sport/leisure injuries treated in hospital emergency rooms in a large metropolitan city. They classified non-contact sports injuries by sport, resulting in the following data set:

Sport
Soccer
Basketball
Others
Basketball
Touch Football
Others
Touch Football
Volleyball
Baseball/softball
⋮

Question: How to summarize this data set?

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.9

Notes

Frequency Table

- A **frequency table** for **categorical data** is a table that displays the possible categories along with the associated **frequencies** or **relative frequencies**
- The **frequency** for a particular category is the number of times the category appears in the data set
- The **relative frequency** for a particular category is the fraction or proportion of the time that the category appears in the data set. It is calculated as:

$$\text{relative frequency} =$$

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.10

Notes

Frequencies and Relative Frequencies

```
> table(sport)
sport
Baseball/softball      Basketball      Bicycling      Jogging/running
      11              19              11              11
      Others      Soccer      Touch Football      Volleyball
      47              24              38              17

> table(sport) / dim(sport)[1]
sport
Baseball/softball      Basketball      Bicycling      Jogging/running
      0.06179775      0.10674157      0.06179775      0.06179775
      Others      Soccer      Touch Football      Volleyball
      0.26404494      0.13483146      0.21348315      0.09550562
```

How could we visualize these information?
⇒ Making a **bar chart** and/or a **pie chart**

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

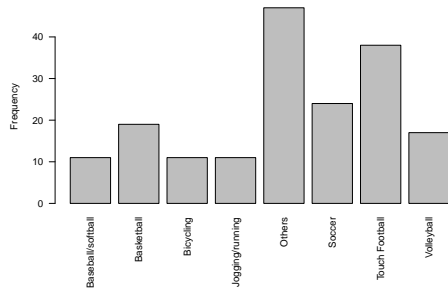
Summarizing Numerical Data

2.11

Notes

Bar Charts

A **bar chart** draws a bar with a height proportional to the count in the table:



Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

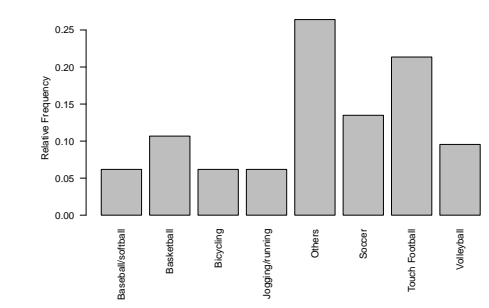
Summarizing Categorical Data

Summarizing Numerical Data

2.12

Notes

Bar Charts cont'd



Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.13

Notes

Pie Charts



Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

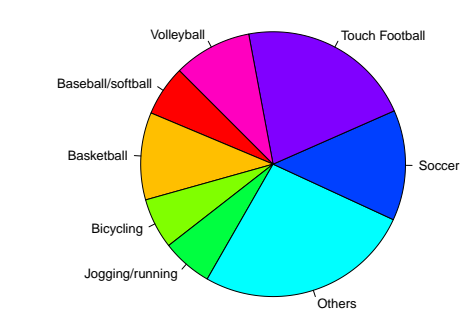
Summarizing Categorical Data

Summarizing Numerical Data

2.14

Notes

Pie Charts cont'd



Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.15

Notes

Bar Charts vs. Pie Charts

Discussion: Which one you prefer to visualize categorical variables. Why?

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

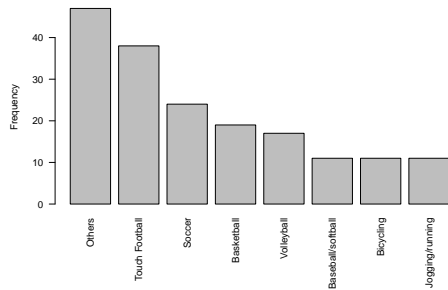
Summarizing Categorical Data

Summarizing Numerical Data

2.16

Notes

A Good Bar Chart



Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

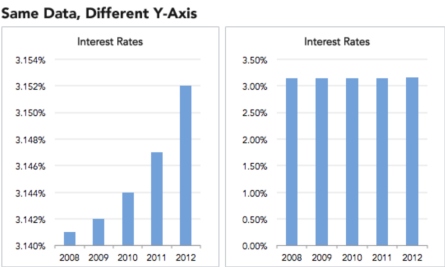
Summarizing Categorical Data

Summarizing Numerical Data

2.17

Notes

A (Potential) Misleading Bar Chart



Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.18

Notes

Example: O'Hare Airport Flight Data



	carrier	origin
1	UA	EW
2	AA	LGA
3	AA	LGA
4	AA	LGA
5	UA	LGA
6	UA	EW

In this example, we have two categorical variables, `carrier` and `origin`, respectively. How to summarize/visualize this dataset?

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

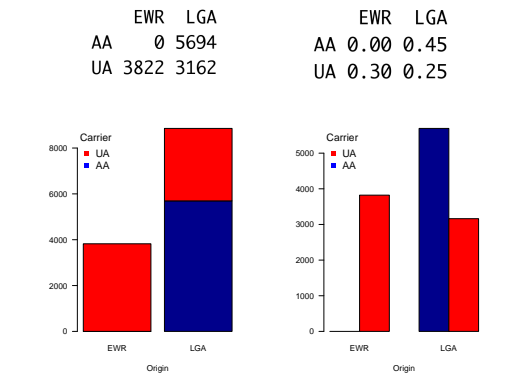
Summarizing Categorical Data

Summarizing Numerical Data

2.19

Notes

ORD Flight Data Cont'd



Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.20

Notes

Summarizing Numerical Variables

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.21

Notes

Example: Murder arrests (per 100,000) in US States in 1973

Data: 13.2, 10.0, 8.1, 8.8, 9.0, 7.9, 3.3, 5.9, 15.4, 17.4, 5.3, 2.6, 10.4, 7.2, 2.2, 6.0, 9.7, 15.4, 2.1, 11.3, 4.4, 12.1, 2.7, 16.1, 9.0, 6.0, 4.3, 12.2, 2.1, 7.4, 11.4, 11.1, 13.0, 0.8, 7.3, 6.6, 4.9, 6.3, 3.4, 14.4, 3.8, 13.2, 12.7, 3.2, 2.2, 8.5, 4.0, 5.7, 2.6, 6.8.

Question: How to graphically summarize this data set?

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

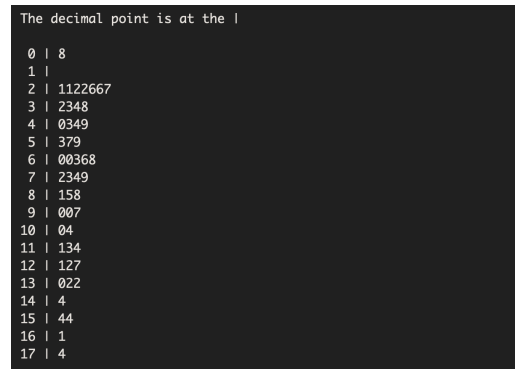
Summarizing Categorical Data

Summarizing Numerical Data

2.22

Notes

Stem-and-Leaf Plot



Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

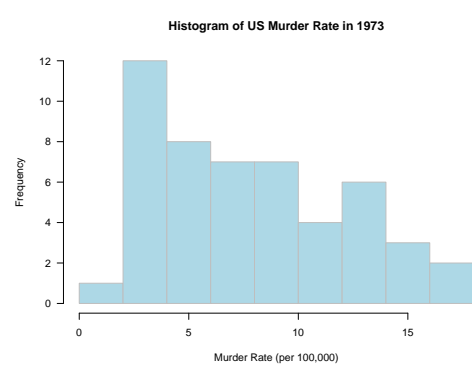
Summarizing Categorical Data

Summarizing Numerical Data

2.23

Notes

Histogram



Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

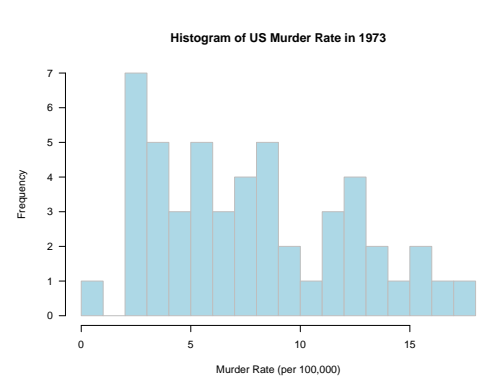
Summarizing Categorical Data

Summarizing Numerical Data

2.24

Notes

Histogram



Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

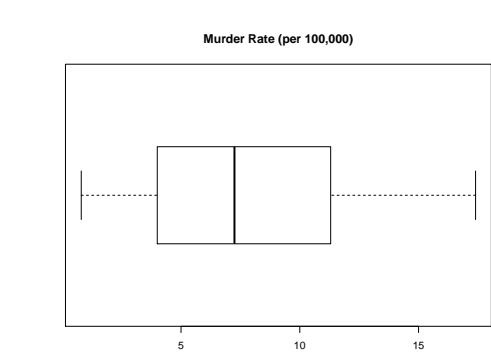
Summarizing Categorical Data

Summarizing Numerical Data

2.25

Notes

Box-and-Whisker Plot



Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

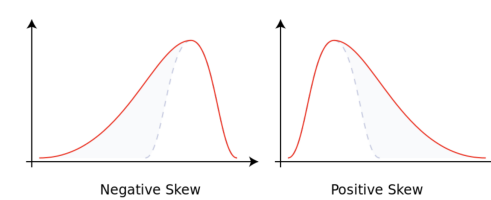
Summarizing Categorical Data

Summarizing Numerical Data

2.26

Notes

Shape of Distributions



Source: [Skewness - Wikipedia](#)

In the rest of the class, we will talk about how to summarize a numerical variable in terms of its **center** and **spread**

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.27

Notes

Measures of Center

- A **measure of center** attempts to report a “typical” value for the variable
- When a measure of center is calculated with **sample data** it is a **statistic**
- When a measure of center is calculated with popular (e.g., census data) it is a **parameter**
- **Measures:** Mean, Median, Mode

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.28

Notes

Mean

- The **population mean**, denoted by μ_X , is the sum of all the population values ($\{X_1, \dots, X_N\}$) divided by the size of the population (N). That is,
- $$\mu_X = \frac{\sum_{i=1}^N X_i}{N}$$
- The **sample mean**, denoted by \bar{X} is the sum of all the sample values ($\{X_1, \dots, X_n\}$) divided by the sample size (n). That is,

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.29

Notes

Median

- The **median** is the value separating the higher half from the lower half of a data sample
- How to compute the median:** Order the n observations in a data set from smallest to largest, then
- Median = $\begin{cases} \text{the single middle value,} & n \text{ odd} \\ \text{the average of the middle two values,} & n \text{ even} \end{cases}$

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.30

Notes

Mode

The **mode** is the value of the observation that appears most frequently

How to compute the mode(s): Order the observations in a data set from smallest to largest, then find the number that is repeated more often than any other

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

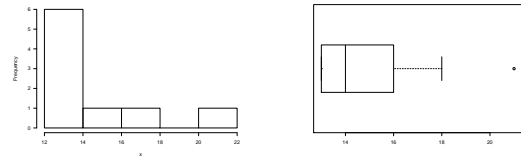
2.31

Notes

Example

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Plot this “data set” and describe the shape of the distribution



Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.32

Notes

Example cont'd

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^9 \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13}{9} = 15$$

- Find the sample median
 - Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21
 - Compute the sample size n and identify (or compute) the median value
 - $n = 9 \Rightarrow$ the median is the 5th number, which is 14

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.33

Notes

Example cont'd

- Find the mode
 - Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21
 - We have 3 13 and 2 14 \Rightarrow 13 is the mode

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.34

Notes

Example: Resistant (Robust) Statistics

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

- Find the sample mean
$$\bar{X} = \sum_{i=1}^9 \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 210 + 13}{9} = 36$$
- Find the sample median
 - Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210
 - Compute the sample size n and identify (or compute) the median value
 - $n = 9 \Rightarrow$ the median is the 5th number, which is (still) 14

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.35

Notes

Example cont'd

- Find the mode
 - Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210
 - We have 3 13 and 2 14 \Rightarrow 13 is (still) the mode

What is the take-home message?

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

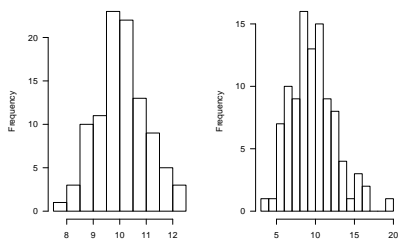
Summarizing Categorical Data

Summarizing Numerical Data

2.36

Notes

Measures of Spread



- Measures: Range, Variance/Standard Deviation, Interquartile range (IQR)

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.37

Notes

Range

The range of a dataset is the difference between the largest and smallest values

Range = Largest Value – Smallest Value

- Compute the range of the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13
- Compute the range of the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

Question: Is Range a robust statistic?

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.38

Notes

Standard Deviation/Variance

- The sample standard deviation (variance), denoted by s (s^2), is a measure of the amount of variation of data. s (s^2) can be used as the estimate of the population standard deviation (variance), denoted by σ (σ^2)
- s is calculated in the following way:
 - Calculate the sample mean \bar{X}
 - Calculate the deviation (from the sample mean) for each observation (i.e., $X_i - \bar{X}$, $i = 1, \dots, n$)
 - Square each deviation and add them (i.e., $\sum_{i=1}^n (X_i - \bar{X})^2$)
 - Divide by $n - 1$ and take the square root, that is,

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.39

Notes

Example

- Compute s of the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13
- Compute s of the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

Question: Is standard deviation a robust statistic?

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.40

Notes

Interquartile range (IQR)

- $IQR = Q_3 - Q_1$, where Q_1 is the Lower Quartile (the median of the lower half of the data) and Q_3 is the Upper Quartile (the median of the upper half of the data)
- Compute the IQR of the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13
- Compute the IQR of the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

Question: Is IQR a robust statistic?

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.41

Notes

Summary

In this lecture, we learned

- Sampling Techniques
- Summarizing Categorical Data
- Summarizing Numerical Data

In next lecture we will learn

- How to construct a boxplot
- How to visualize numerical + categorical variables and numerical + numerical variables
- How to visualize time series, cross-sectional, and spatio-temporal Data sets

Data Summary/Visualization I

CLEMSON UNIVERSITY

Sampling Techniques

Summarizing Categorical Data

Summarizing Numerical Data

2.42

Notes
