Lecture 3

Descriptive Statistics I

Text: Chapter III

STAT 8010 Statistical Methods I
August 26, 2019

Whitney Huang
Clemson University

Notes

Agenda

Descriptive Statistics I
CLEMSON UNIVERSITY
Review of Last Class
Summarizing Categorical Data
Summarizing Numerical Data

1. **Review of Last Class**
2. **Summarizing Categorical Data**
3. **Summarizing Numerical Data**

Notes

Last Lecture

Descriptive Statistics I
CLEMSON UNIVERSITY
Review of Last Class
Summarizing Categorical Data
Summarizing Numerical Data

- Stating the problem, identifying the variable(s) of interest, and gathering data
  - Types of variables
  - 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
- Summarizing the data
- Analyzing the data
- Reporting and interpreting the results

Notes

Example

The paper "PROFILE OF SPORT/LEISURE INJURIES TREATED AT EMERGENCY ROOMS OF URBAN HOSPITALS." by Pelletier, R. L., G. Anderson, and R. M. Stark, 1991 (Link to the abstract https://europepmc.org/abstract/med/1647867) 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 (**Link**: https://whitneyhuang83.github.io/sport.txt):

Sport
Soccer
Basketball
Basketball
Basketball
⋮

Question: How to summarize this data set?

Notes

Frequency Distribution

- A frequency distribution 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:

$$\text{relative frequency} = \frac{\text{frequency}}{\text{number of observations}}$$

Notes

Frequencies and Relative Frequencies

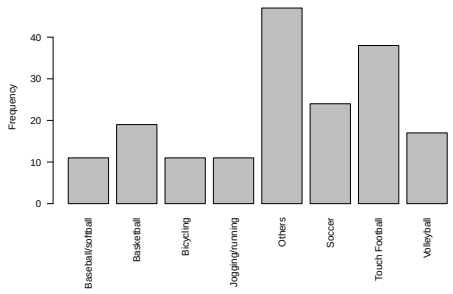
```
> table(sport)
sport
Baseball/softball        Basketball         Bicycling   Jogging/running
               11                19                11                11
           Others            Soccer   Touch Football        Volleyball
               47                24                38       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

Can we plot these information? ⇒ **Bar charts** and **Pie charts**

Notes

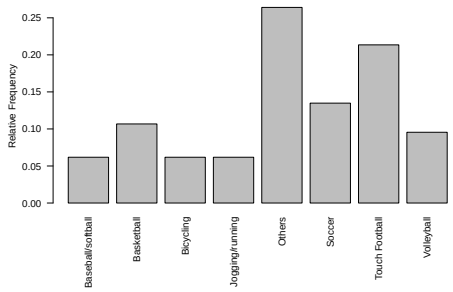
Bar Charts

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



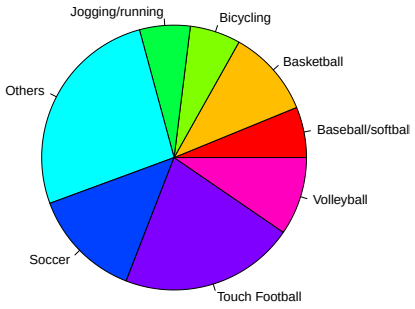
Notes

Bar Charts cont'd



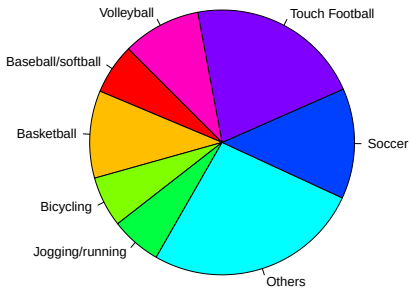
Notes

Pie Charts



Notes

Pie Charts cont'd



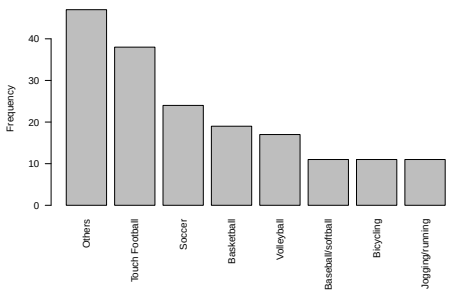
Notes

Bar Charts vs. Pie Charts

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

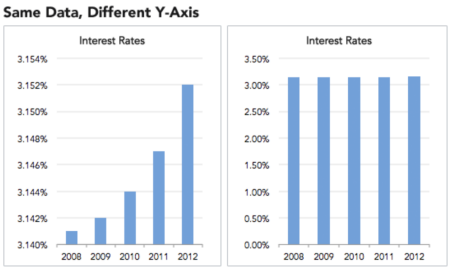
Notes

A Good Bar chart



Notes

A Bad Bar chart: Truncated Bar Chart



Notes

Example: Max Heart Rate and Age

Suppose we have 15 people of varying ages are tested for their maximum heart rate (bpm)

Age	18	23	25	35	65	54	34	56	72	19	23	42	18	39	37
MaxHeartRate	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178

Link to this dataset: http://whitneyhuang83.github.io/maxHeartRate.csv

- How many variables do we have in this data set? What are the variable types?
- How to summarize these variables?

Notes

Numerical Summaries of Quantitative Variables

- Mean: the average/expected value of a set of numbers
  - Population mean: μ_x
  - Sample mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Variance: measures how far a set of numbers is spread out
  - Population variance: $\sigma_x^2 = \frac{\sum_{i=1}^{N} (x_i - \mu_x)^2}{N}$
  - Sample variance: $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
- Mode: the value that appears most often in a set of numbers
- Range: the largest value $-$ the smallest value in a set of numbers

Notes

Example

Suppose we have the data set 1, 2, 3, 4, and 5. Find the mean of the data. Also compute variance in 2 ways (one assuming that this is a sample, the other assuming that this represents the entirety of the population)

Solution.

- Mean: $\bar{x} = \frac{1+2+3+4+5}{5} = 3$
- Sample variance: $s^2 = \frac{\sum_{i=1}^{5} (x_i - 3)^2}{5-1} = \frac{10}{4} = 2.5$
- Population variance: $\sigma^2 = \frac{\sum_{i=1}^{5} (x_i - 3)^2}{5} = \frac{10}{5} = 2$

Notes

Notes