Lecture 3

Data Summary/Visualization II

Text: Chapter 3

STAT 8010 Statistical Methods I
August 27, 2020

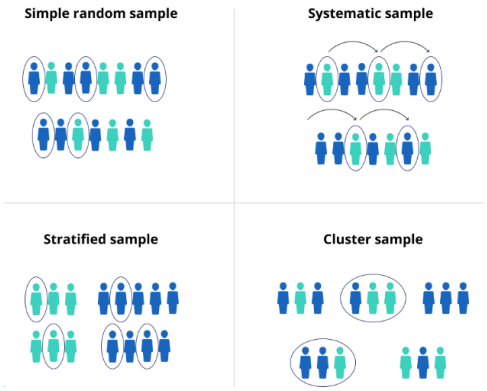
Whitney Huang
Clemson University

Notes

Agenda

Notes

Last Lecture: Sampling Techniques

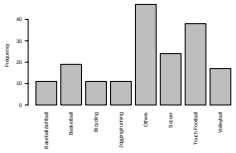


Source: `https://www.scribbr.com/methodology/sampling-methods/`

Notes

Last Lecture: Summarizing Categorical Variables

```
> 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
```



Notes

Summarizing Numerical Variables

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?

Notes

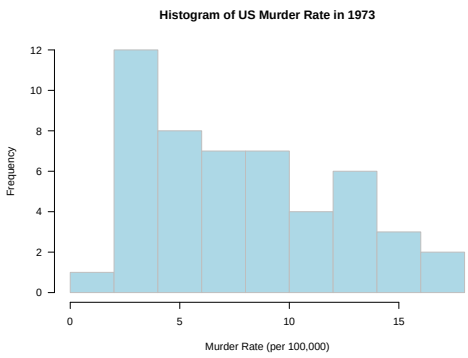
Stem-and-Leaf Plot

```
The decimal point is at the |

 0 | 8
 1 |
 2 | 1122667
 3 | 2348
 4 | 0349
 5 | 379
 6 | 00368
 7 | 2349
 8 | 158
 9 | 007
10 | 04
11 | 134
12 | 127
13 | 022
14 | 4
15 | 44
16 | 1
17 | 4
```

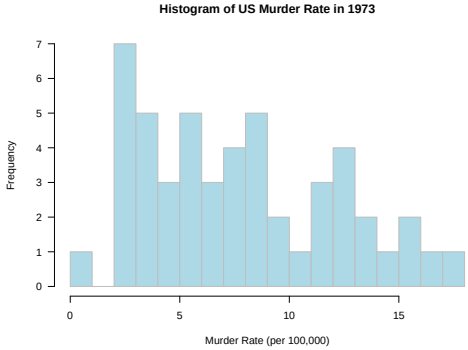

Notes

Histogram



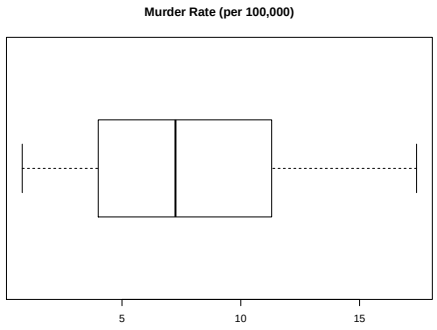
Notes

Histogram



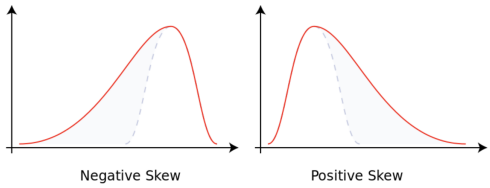
Notes

Box-and-Whisker Plot



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

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

Notes

Mean

- The population mean, denoted by μ_X, is the sum of all the population values ($\{X_i, \cdots, X_N\}$) 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 ($\{X_1, \cdots, X_n\}$) divided by the sample size (n). That is,

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

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

$$\text{Median} = \begin{cases} \text{the single middle value,} & \text{n odd} \\ \text{the average of the middle two values,} & \text{n even} \end{cases}$$

Notes

Mode

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

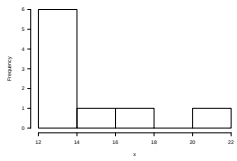
How to compute the mode(s): Order the observations in a data set from smallest to largest, then find the number that is repeated more often than any other

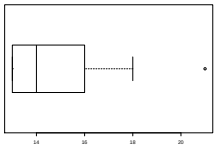
Notes

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





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

Notes

Example cont’d

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

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
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210
  2. Compute the sample size n and identify (or compute) the median value
  3. $n = 9 \Rightarrow$ the median is the 5th number, which is (still) 14

Notes

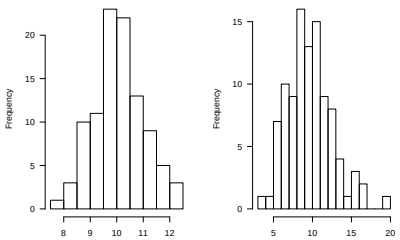
Example cont'd

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

What is the take-home message?

Notes

Measures of Spread



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

Notes

Range

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

$$\boxed{\text{Range} = \text{Largest Value} - \text{Smallest Value}}$$

- 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?

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 (varaince), denoted by σ (σ^2)
- s is calculated in the following way:
  1. Calculate the sample mean $\bar{X}$
  2. Calculate the deviation (from the sample mean) for each observation (i.e., $X_i - \bar{X}, \quad i = 1, , \cdots, n$)
  3. Square each deviation and add them (i.e., $\sum_{i=1}^{n}(X_i - \bar{X})^2$)
  4. Divide by $n - 1$ and take the square root, that is,

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

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?

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?

Notes

Percentiles, Quartiles, and Boxplots

Notes

Percentiles

- The p_{th} percentile is a value such that at least p% of the data set is less than or equal to this value [An Example]
- Calculation of percentiles using the indexing method:
  1. Sort the set of numbers in an increasing order
  2. For the p_{th} percentile, compute the index $i = \frac{np}{100}$ where n is the sample size
  3. If i is an integer then p_{th} percentile is the average of i_{th} value and $(i+1)_{\text{th}}$ value, otherwise take the $(i+1)_{\text{th}}$ value
- Quartiles:
  1. $Q1$: first quartile (25_{th} percentile)
  2. M ($Q2$): median (second quartile, 50_{th} percentile)
  3. $Q3$: third quartile (75_{th} percentile)
  4. Interquartile range or IQR: $Q3 - Q1$

Notes

Example

Find Q_1, M, Q_3 and IQR of the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13 using the indexing method

1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21
2. Find the sample size n and compute the indices for $p = 25, 50, 75$
3. $n = 9 \Rightarrow$ the indices are $3, 5, 7 \Rightarrow Q_1 = 13$, $M = 14$, $Q_3 = 16$
4. IQR $= Q_3 - Q_1 = 16 - 13 = 3$

Notes

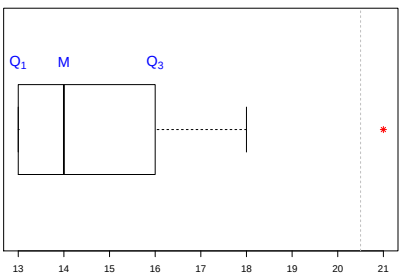
Steps to Making a Boxplot

1. Find Q_1, M, Q_3 and draw a box from Q_1 to Q_3. Add a vertical line inside the box at M
2. Compute the value of Lower Fence (LF) $= Q1 - 1.5$IQR and the Upper Fence (UF) $= Q3 + 1.5$IQR. Find the largest value $\leq$ UF and the smallest value $\geq$ LF. Draw whiskers go from Q_1, Q_3 to these two values
3. Plot the individual outlier(s) (i.e., the values either $>$ UF or $<$ LF)

Notes

Bopxplot

- **Ordered data values:** 13, 13, 13, 13, 14, 14, 16, 18, 21
- **IQR** $16 - 13 = 3 \Rightarrow$ LF = $13 - 1.5 \times 3 = 8.5$; UF = $16 + 1.5 \times 3 = 20.5$



Notes

Example

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13, 9, 27, 18, 25, 20, 6

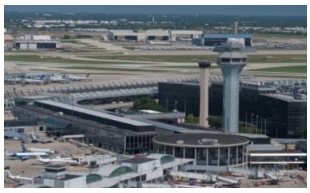
- Find the 35th percentile
  1. Sort the data:
     6, 9, 13, 13, 13, 13, 14, 14, 16, 18, 18, 20, 21, 25, 27
  2. Compute the index value $i = \frac{35 \times 15}{100} = 5.25 \Rightarrow$ the 35th percentile is 13
- Find the 65th percentile
  1. Sort the data:
     6, 9, 13, 13, 13, 13, 14, 14, 16, 18, 18, 20, 21, 25, 27
  2. Compute the index value $i = \frac{65 \times 15}{100} = 9.75 \Rightarrow$ the 65th percentile is 18

Notes

Visualizing two variables simultaneously

Notes

Example: O'Hare Airport Flight Data



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

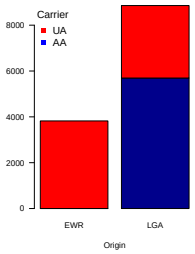
In this example, we have two categorical variables, `carrier` and `origin`, respectively. How to summarize/visualize this dataset?

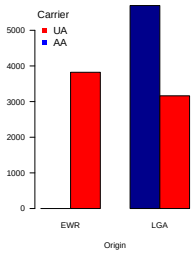
Notes

ORD Flight Data Cont'd

```
      EWR  LGA
AA      0 5694
UA   3822 3162
```

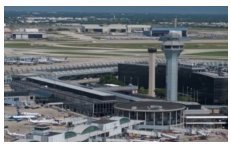
```
      EWR  LGA
AA 0.00 0.45
UA 0.30 0.25
```





Notes

ORD Fligts Data Cont'd

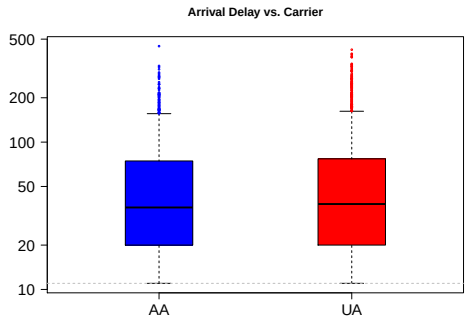


```
carrier origin arr_delay
     UA    EWR        12
     AA    LGA         8
     AA    LGA        14
     AA    LGA         4
     UA    LGA        20
     UA    EWR        21
```

In this example, we have two categorical variables, `carrier`, `origin` and a numerical variable `arr_delay`, respectively. How to visualize, for example, `arr_delay` vs. `carrier`?

Notes

ORD Example: Arrival Delay vs. Air Carrier



Notes

Example: Max Heart Rate and Age

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

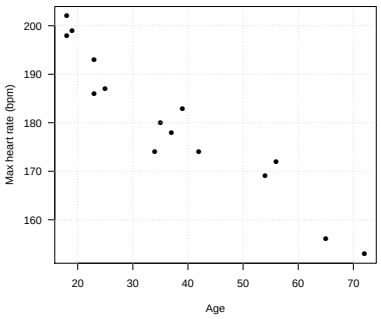
Age	18	23	25	35	65	54	34	56	72	19	23	42	18	39	37
MHR	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178

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

Notes

Scatterplot

A scatterplot is a useful tool to graphically display the relationship between two numerical variables. Each dot on the scatterplot represents one observation from the data

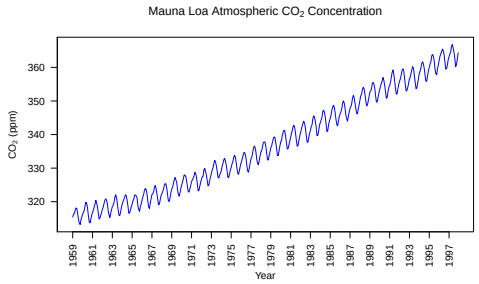


Notes

Visualizing Time Series, Cross-Sectional, and Spatio-Temporal Data sets

Notes

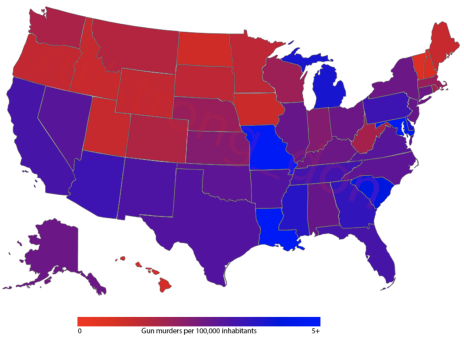
Visualizing Time Series Data



CLEMSON UNIVERSITY

Notes

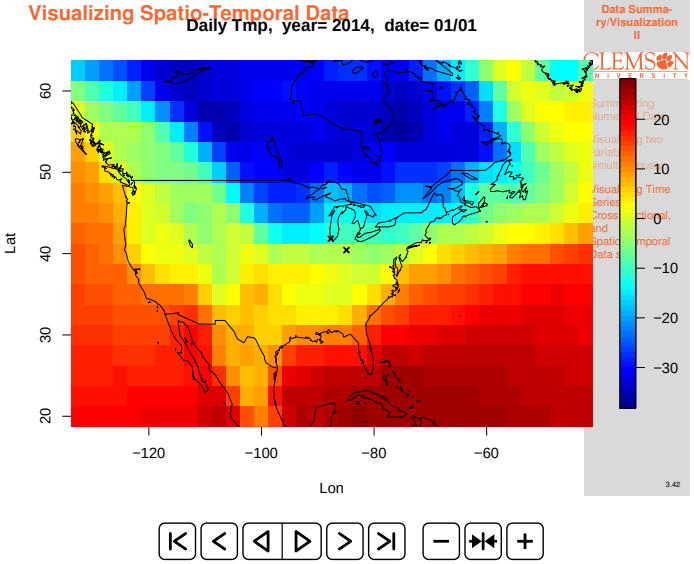
Visualizing Cross-Sectional Data



CLEMSON UNIVERSITY

Notes

Visualizing Spatio-Temporal Data



CLEMSON UNIVERSITY

Notes

Summary

In this lecture, we learned

- How to summarize numerical variable
- How to visualize two variables simultaneously
- How to visualize time series, cross-sectional, spatio-temporal data sets

We will talk about Probability in the next few weeks

Notes

Notes

Notes