Lecture 4

Data Summary/Visualization III

Text: Chapter 3

STAT 8010 Statistical Methods I
September 1, 2020

Whitney Huang
Clemson University

Notes

Agenda

1. Percentiles and Boxplots
2. Visualizing two variables simultaneously
3. Visualizing Time Series, Cross-Sectional, and Spatio-Temporal Data sets

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. $\text{IQR} = Q_3 - Q_1 = 16 - 13 = 3$

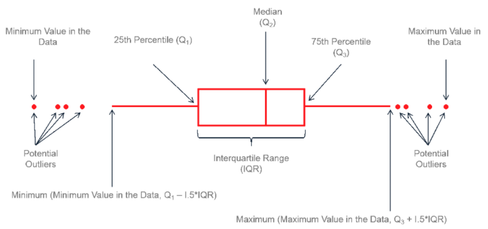
Notes

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) $= Q1 - 1.5\text{IQR}$ and the Upper Fence (UF) $= Q3 + 1.5\text{IQR}$. 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

Boxplot Anatomy

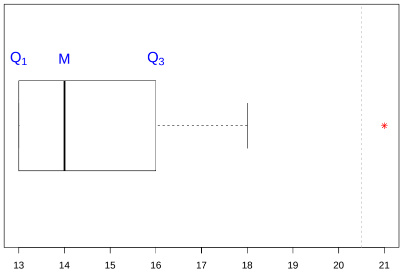

Minimum Value in the Data
25th Percentile (Q_1)
Median (Q_2)
75th Percentile (Q_3)
Maximum Value in the Data
Potential Outliers
Interquartile Range (IQR)
Potential Outliers
Minimum (Minimum Value in the Data, Q_1 – 1.5*IQR)
Maximum (Maximum Value in the Data, Q_3 + 1.5*IQR)

Source: `https://https://www.leansigmacorporation.com/box-plot-with-minitab/`

Notes

Example

- **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

Another 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 $i = \frac{35\times15}{100} = 5.25 \Rightarrow$ 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 $i = \frac{65\times15}{100} = 9.75 \Rightarrow$ 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
```

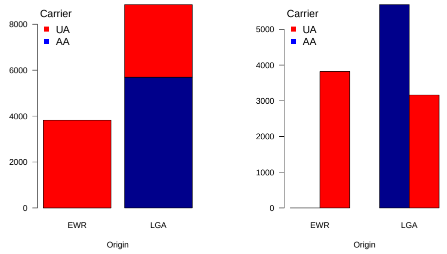
In this example, we have two categorical variables, `carrier` and `origin`, respectively. How to summarize/visualize this dataset?

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

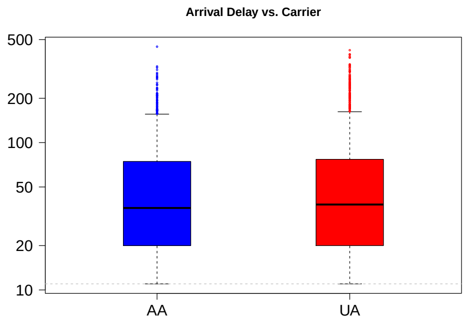
ORD Fligts Data Cont'd

```
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)

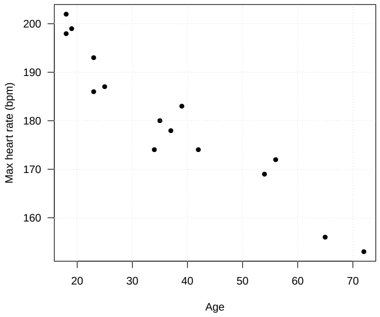
Age	18	23	25	35	65	54	34	56	72	19	23	42	18	39	37
MHR	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178

- 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

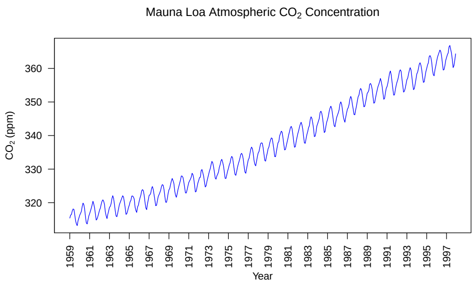


Notes

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

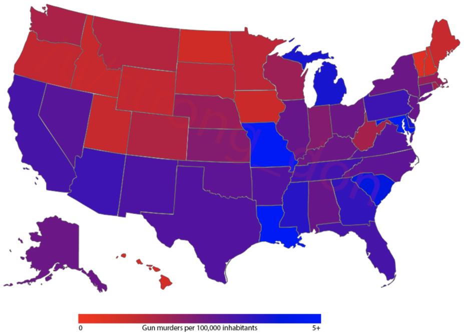
Notes

Visualizing Time Series Data



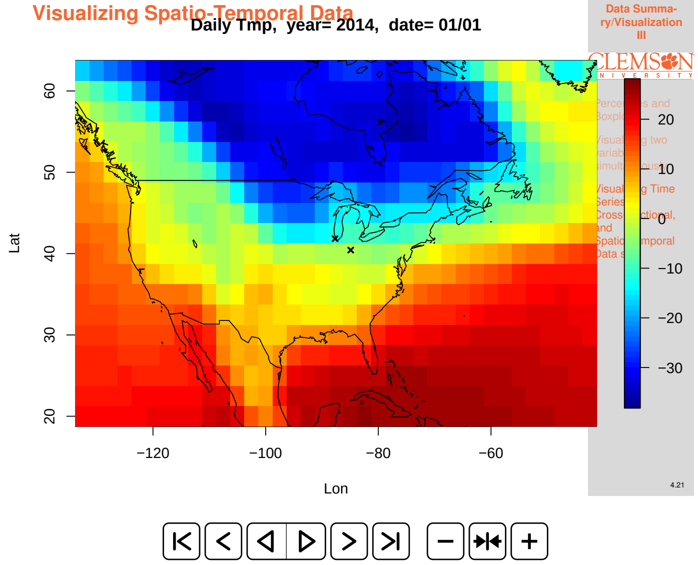
Notes

Visualizing Cross-Sectional Data



Notes

Visualizing Spatio-Temporal Data



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