Data Summaries

Grinnell College

January 31, 2024

Review

What is a **distribution**?

Types of bar plots

How might we determine when variables associated?

Scatterplots

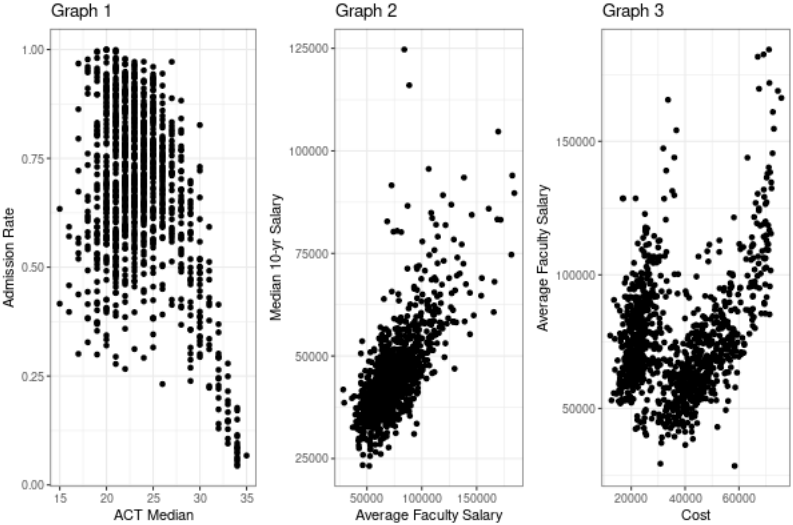
Scatterplots show relationships between two quantitative variables. When describing an association, we should address the following:

1. **Form** – what type of trend or pattern exists (linear, non-linear, exponential, etc.,)
2. **Strength** – how closely do the data adhere to a trend or pattern (i.e., strong, moderate, weak)
3. **Direction** – how the values of one variable relate to the values of another variable (i.e., positive, negative)

Note: For some non-linear associations you may not be able to provide a single direction

Scatterplots

How would you describe the following associations?



Transformations

maybe this

spread variability?

maybe have this here, 68, 95, 99 rule

Percentiles

A **percentile** α is a number such that α% of our (quantitative) observations fall below this number when ranked from smallest to largest

The *median*, for example, is the 50th percentile. Other notable percentiles include:

1. Minimum
2. 25th percentile or **first quartile** (Q_1)
3. 75th percentile or **third quartile** (Q_3)
4. Maximum

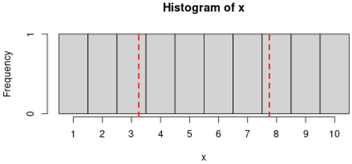
Along with the median, these numbers make up the *five-number summary* for describing data

IQR

The **interquartile range** or **IQR** is the value of $Q_3 - Q_1$, giving the breadth of the middle 50% of the observed data

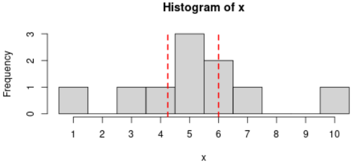
$x = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

- $x_{\{25\}} = 3.25$, $x_{\{75\}} = 7.75$
- $IQR = 4.5$

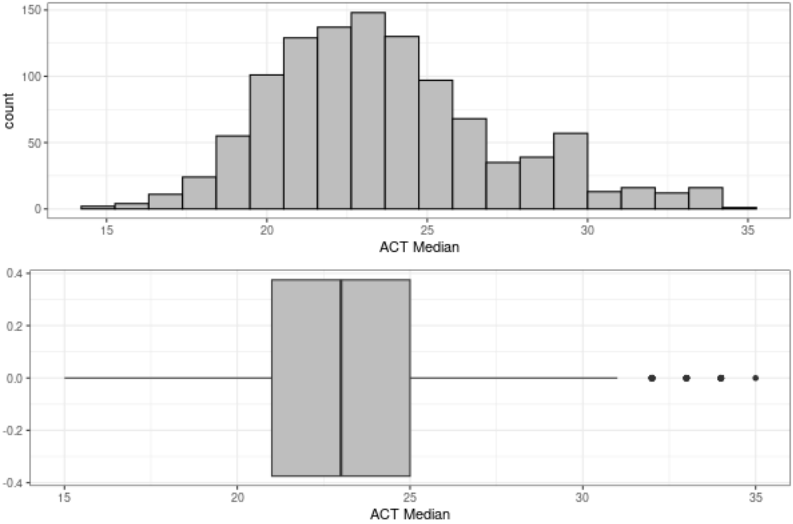


$x = \{1, 3, 4, 5, 5, 5, 6, 6, 7, 10\}$

- $x_{\{25\}} = 4.25$, $x_{\{75\}} = 6$
- $IQR = 1.75$

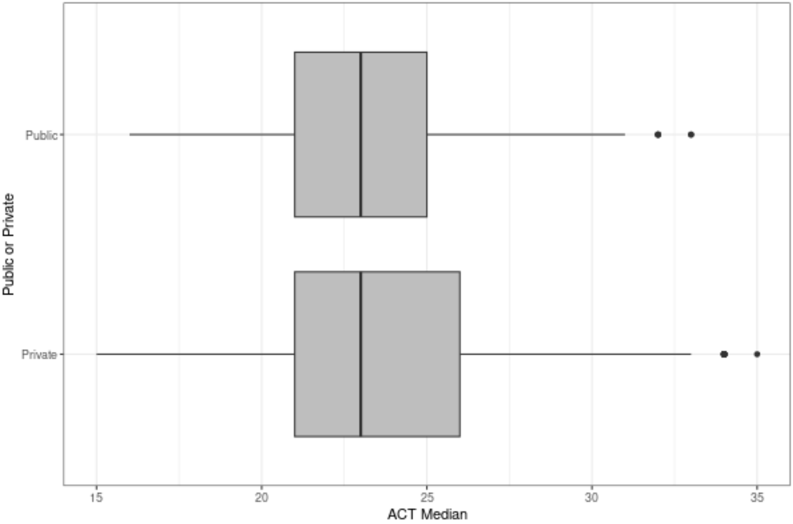


Box plots

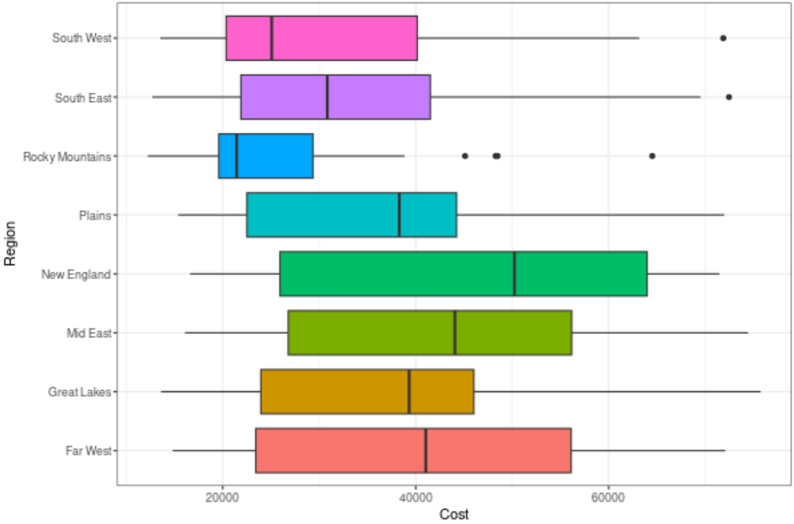


count
150
100
50
0
15
20
25
30
35
ACT Median
0.4
0.2
0.0
-0.2
-0.4
15
20
25
30
35
ACT Median

Box plots



Box plots


South West
South East
Rocky Mountains
Plains
New England
Mid East
Great Lakes
Far West
Region
20000
40000
60000
Cost

Review

- Why summarize?
- Identify appropriate univariate plots for each variable type and use to describe distribution
  - Shape, center, spread
  - Counts and frequency
- Identify appropriate bivariate plots to describe possible associations
  - Scatterplots – form, strength, and direction
  - Bar charts – stacked, dodged/clustered, conditional
  - Box plots – five number summary