Descriptive Statistics

Part 2 - Quantitative variables

Ryan Miller

Grinnell College
Statistics

Introduction

Data visualizations provide a qualitative assessment of the distribution of a variable, or the relationship between two variables:

- Distributions of one variable (univariate graphs):
  - One categorical variable - bar chart
  - One quantitative variable - histogram or box plot
- Relationships between two variables (bivariate graphs):
  - Two categorical variables - stacked, clustered, or conditional bar chart
  - Two quantitative variables - scatter plot
  - One categorical and one quantitative variable - side-by-side box plots or histograms

Introduction

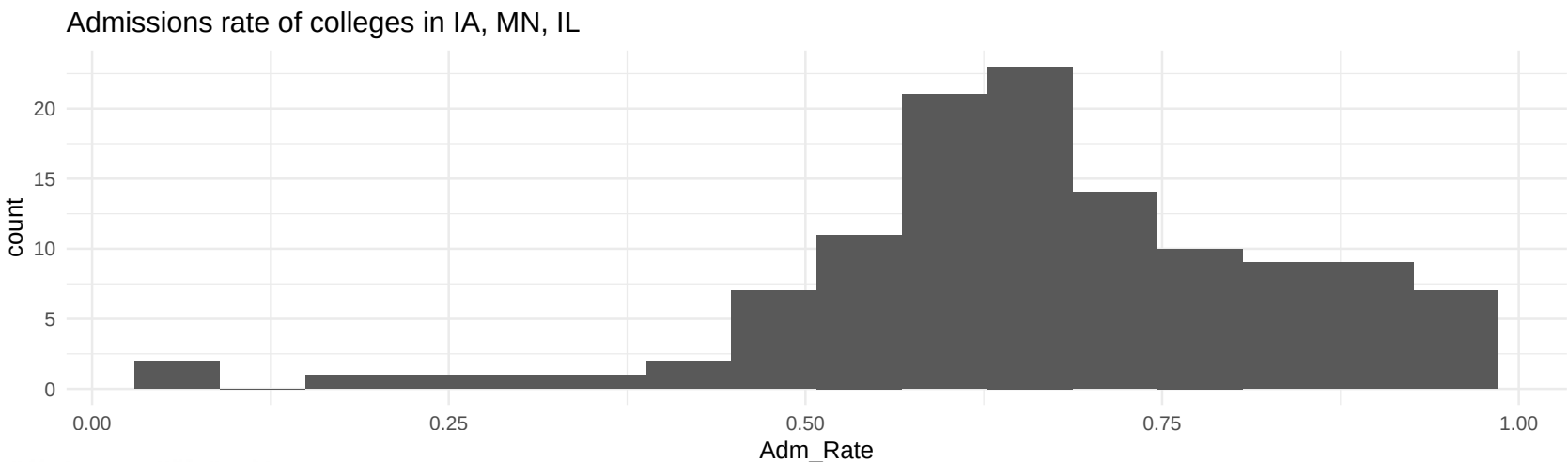
Descriptive statistics provide a quantitative summary of a variable's distribution or a relationship between variables:

- So far, we've covered scenarios involving one or two categorical variables
  - One categorical variable - one-way frequency table or proportions
  - Two categorical variables - two-way frequency tables, conditional proportions, risk difference, relative risk, and odds ratio
- Today we'll cover situations involving one quantitative variable, and those involving one categorical and one categorical variable

Describing the distribution of a quantitative variable

Recall that we should consider four aspects of a quantitative variable's distribution:

1. **Shape** - is the distribution symmetric or skewed? is it bell-shaped?
2. **Center** - where is the distribution centered at?
3. **Spread** - how much do values of the variable tend to vary?
4. **Unusual points** - are there any outliers? excessive zeros or anomalies?



Describing a quantitative variable's "center"

We have two summarize a variable's center:

- **Mean** - the arithmetic average of a variable, if we have n observations the mean of variable "X" is given by: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **Median** - the middle value if the data were arranged in ascending order

The median is often called a **robust** measure of center because it tends not be influenced by outliers. In contrast, the mean tends to be pulled towards outliers.

Describing a quantitative variable's "spread"

Important ways to summarize a variable's spread:

- **Standard deviation** - the average deviation (distance) of individual data-points from the distribution's mean
  - $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$
- **Range** - the difference in the data's maximum and minimum values
- **Interquartile Range (IQR)** - the difference in the 75^{th} and 25^{th} percentiles of the data (also called Q3 and Q1 respectively)

The standard deviation and range are *greatly* influenced by outliers, while the IQR is resistant/robust.

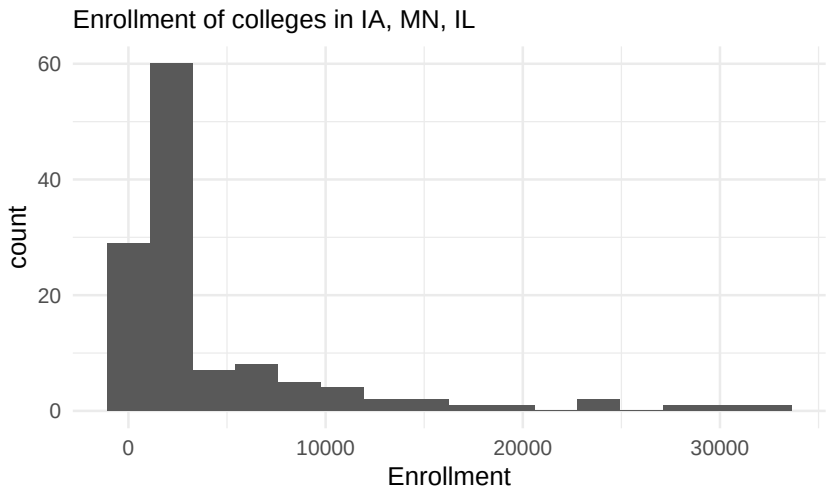
Describing a quantitative variable's "shape" and "outliers"

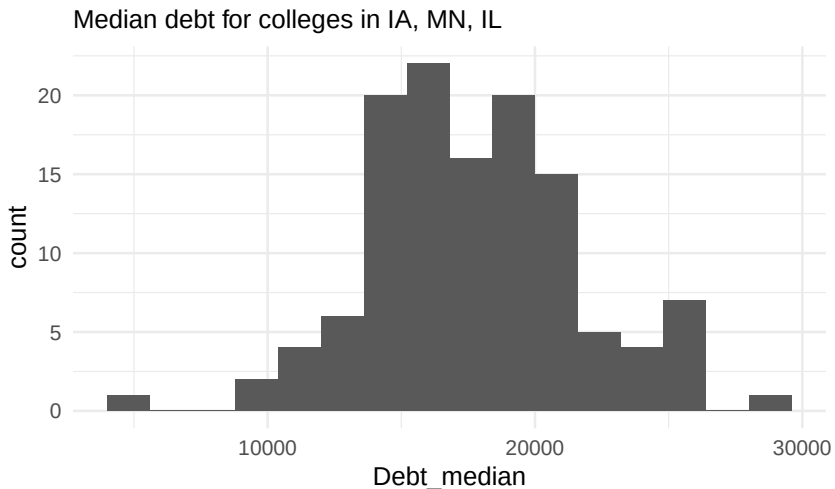
- These are important aspect's of a quantitative variable's distribution that we should consider, however you will not be responsible for describing them numerically
  - Outliers can be expressed as a frequency (ie: there are _ outliers)
  - A simple way to describe skew is the quantity $\frac{\bar{x}-m}{s}$, or the standardized difference between the mean and median

Practice

For each of the following variables (visualized below):

1. Determine an approximate mean and median, and briefly explain how you know which will be larger.
2. Decide whether it's more appropriate to describe spread using standard deviation or IQR.





One quantitative and one categorical variable

- Associations between a categorical and quantitative variable can be assessed using *comparative summaries*
  - The basic idea is to calculate and compare conditional descriptive statistics (ie: describe each group created by the categorical variable separately)
  - Differences across groups indicate the variables are associated

Table 1: Comparative summary statistics for the age (days) of SIDS cases by Sex (GHC of Puget Sound, 1972-1983)

Sex	Min	Q1	Mean	Median	Q3	Max	StDev
F	53	56.0	63.20000	60	60.0	87	13.62718
M	46	77.5	96.45455	81	114.5	175	36.77870

Conclusion

Descriptive statistics are numerical summaries of a distribution or an association. After today, you should understand the following:

- Center - mean (impacted by outliers/skew), median (robust)
- Spread - standard deviation and range (impacted by outliers/skew), IQR (robust)

Association between a categorical and quantitative variable is assessed by calculating separate sets of conditional descriptive statistics

- Differences in means across groups are the most common way to summarize an association