Data Visualization

Ryan Miller

Grinnell College
Statistics

Motivation

Below are some data describing primarily undergraduate colleges with at least 400 students. Do the alumni of private colleges tend to earn more 10 years after graduation?

Name	Private	Salary10yr_median
Tennessee Technological University	Public	40500
Greensboro College	Private	36900
Simpson University	Private	34500
Lubbock Christian University	Private	41800
Loyola University Chicago	Private	54100
Trinity University	Private	54900
University of Kansas	Public	48800
Northwest Missouri State University	Public	40000
Earlham College	Private	35000
Oklahoma Christian University	Private	38500
Texas A & M University-Corpus Christi	Public	43400
Centre College	Private	45500
Aquinas College	Private	37300
California Baptist University	Private	42600
Colorado State University-Pueblo	Public	37500
Pennsylvania State University-Penn State Scranton	Public	50100
University of Wisconsin-Superior	Public	36700

Data presentation

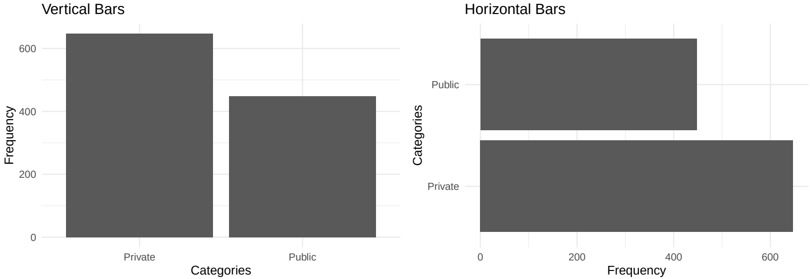
Simply inspecting the raw data is inefficient and rarely useful, better approaches involve:

1. **Data visualization** - graphically displaying the data in ways that make patterns more easily visible
2. **Numerical summaries** - calculating numbers that encapsulate certain aspects of the data

There are many different types of data visualizations and numerical summaries, and choosing the proper one depends upon the *type of variable(s)* as well as the *distribution* of the variable(s).

Distributions (categorical variables)

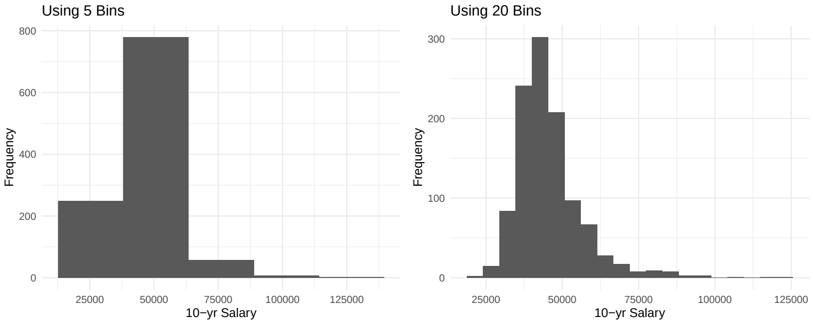
A **distribution** describes how frequently certain values will be observed in a variable across cases



The distribution of a categorical variable can be displayed using a **bar chart**, which shows the frequency of each category present in our data via a position on the x or y axis.

Distributions (quantitative variables)

A **histogram** is a similar visualization used for quantitative variables. Histograms group numeric values into equally spaced intervals known as *bins*, then display the frequencies of data in each bin:



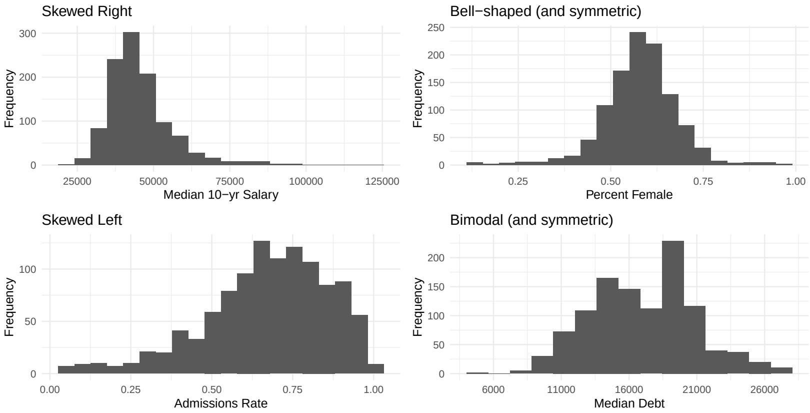
Grinnell College
Statistics

Describing a distribution

- Describing the distribution of a *categorical variable* is straightforward
    - We might highlight the more common and less common categories
- Describing the distribution of a *quantitative variable* is more nuanced, we should address the following:
    - **Shape** - is the distribution symmetric, skewed, bell-shaped, bimodal?
    - **Center** - where are the data centered? (ie: approximate mean or median)
    - **Variability** - how spread out are the data? (ie: range)
    - **Unusual Points** - are there any outliers or excessive zeros?

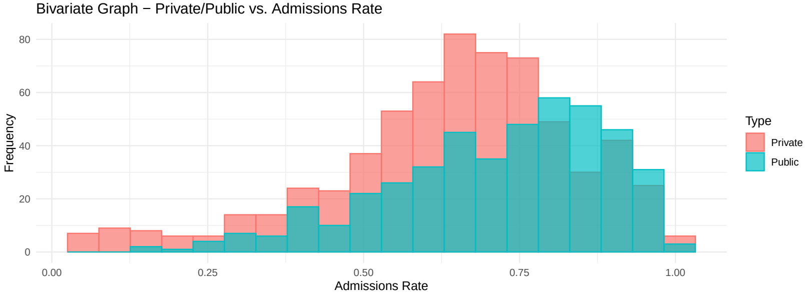
Describing a distribution

Below are a few examples of how we should describe *shape*:



Bivariate graphs

- Our previous examples all were **univariate graphs**, which show the distribution of a single variable
  - **Bivariate graphs** show the relationship between two variables:



Do these variables (admissions rate and type of college) seem related?

Association

- Two variables are **associated** if the value of one variable tells you something about the value of the variable
  - For example, if a college is “public” you’d expect a higher admission rate
  - This is because the *center* of the distribution of admission rates of public colleges is *different* from the center of the distribution for private colleges

Explanatory and response variables

- When discussing an association between two variables we'll sometimes want designate an **explanatory variable** (suspected cause) and a **response variable** (suspected effect)
  - This is usually done via subject-area expertise
    - For example, colleges don't switch from public to private when their acceptance rate reaches a certain point, but being public or private may influence how they judge applicants
- However, seeing that an explanatory and response variable are associated in our data doesn't necessarily confirm a cause-effect relationship
  - We'll discuss criteria for causation later this semester

Variable types

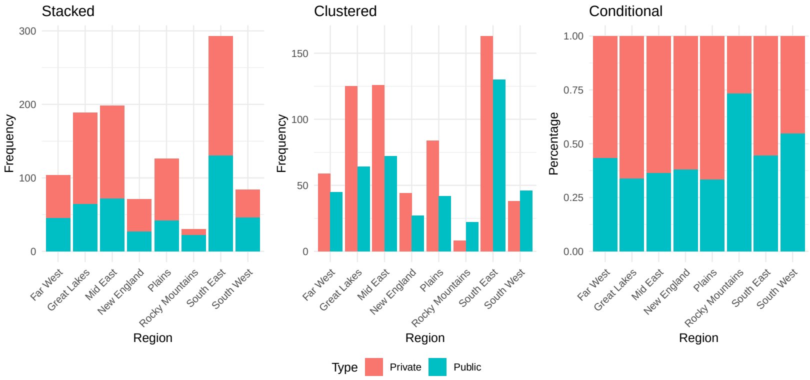
Different combinations of variable types necessitate different types of bivariate graphs:

- 1 categorical and 1 quantitative variable - side-by-side histograms or box plots
- 2 categorical variables - stacked, clustered, or conditional bar charts
- 2 quantitative variables - scatter plots

We'll next see some examples and discuss how we'd use them to identify and describe associations between variables

Bivariate bar charts

Are the variables “Region” and “Type” associated? Which bar chart is most helpful?



Grinnell College
Statistics

Scatter plots

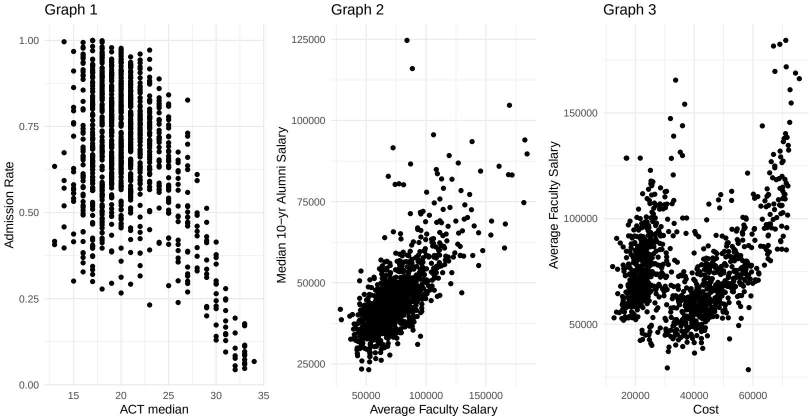
Scatter plots show relationships between two quantitative variables. When describing an association we should address the following:

1. **Form** - what type of trend or pattern exists (ie: linear, non-linear, none)
2. **Strength** - how closely do the data adhere to a trend or pattern (ie: strong, moderate, weak)
3. **Direction** - how the values of one variable relate to the values of the other variable (ie: positive, negative)

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

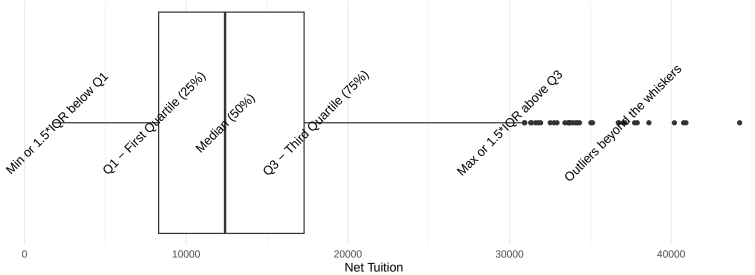
Scatter plots

How would you describe the following associations?



Box plots

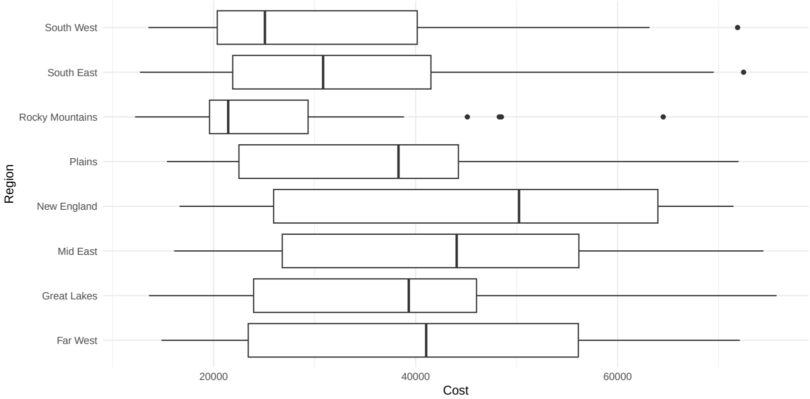
- A **percentile** is a value that a certain proportion of the observed data falls below
  - For example, the 50th percentile is the **median**, while the 90th percentile is larger than 90% of observed data-points
- A **box plot** displays a set of percentiles
  - The IQR (interquartile range) is Q3 (75th percentile) - Q1 (25th percentile)



Grinnell College
Statistics

Side-by-side box plots

Side-by-side box plots tend to be more effective than side-by-side histograms when comparing the distributions of 3 or more groups:



Grinnell College
Statistics

Conclusion

After this lecture and the corresponding labs you should be able to:

1. Identify appropriate univariate graphs for categorical and quantitative variables and use them to describe a variable's distribution.
   - Describe shape, center, spread, and unusual points using a histogram
2. Identify appropriate bivariate graphs for each possible combination of categorical and quantitative variables and use them to describe possible associations.
   - Describe the form, strength, and direction of an association seen in a scatter plot
   - Compare distributions using side-by-side boxplots or histograms, or stacked/clustered/conditional bar charts