Categorical Descriptive Statistics

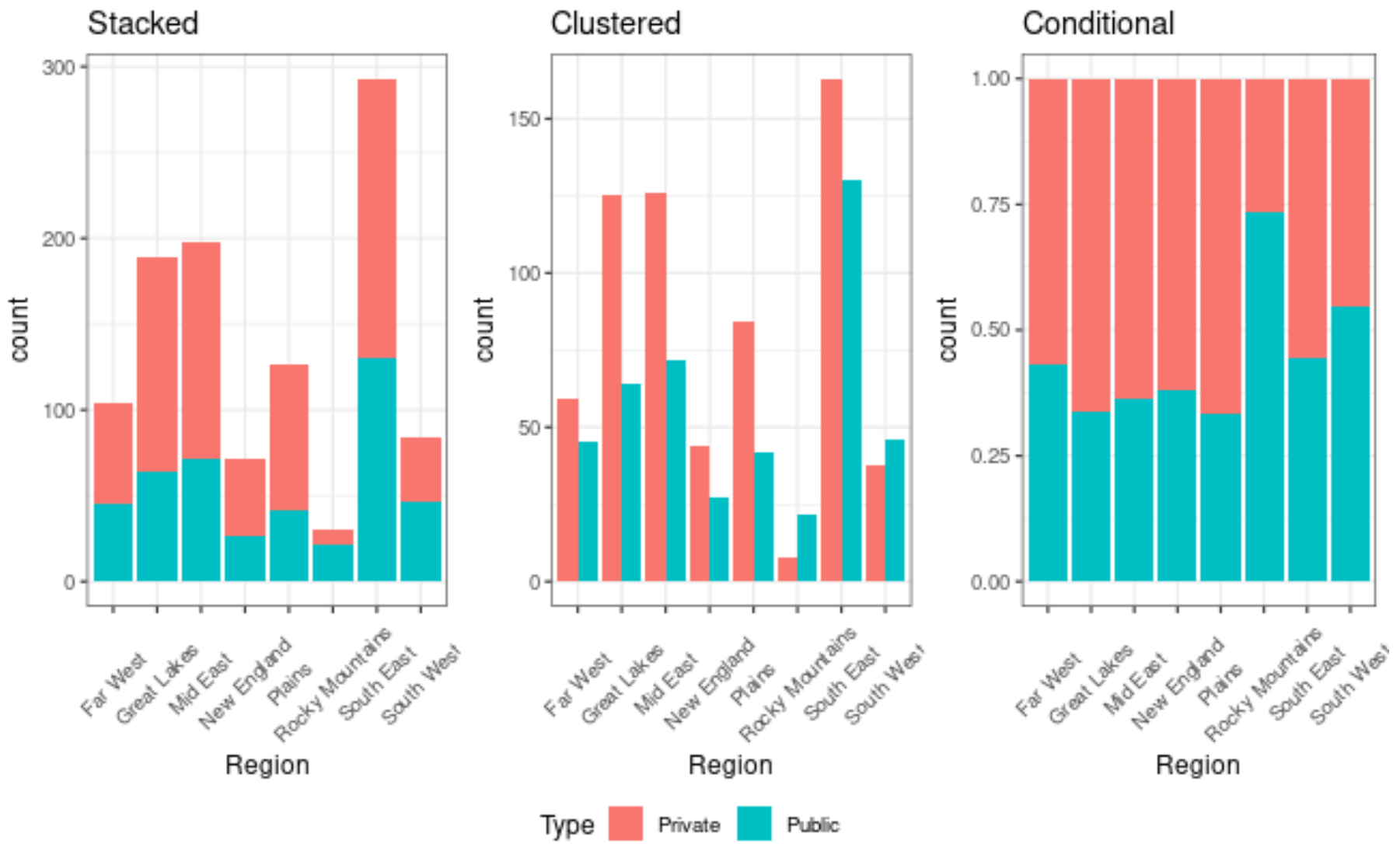
Grinnell College

September 13, 2024

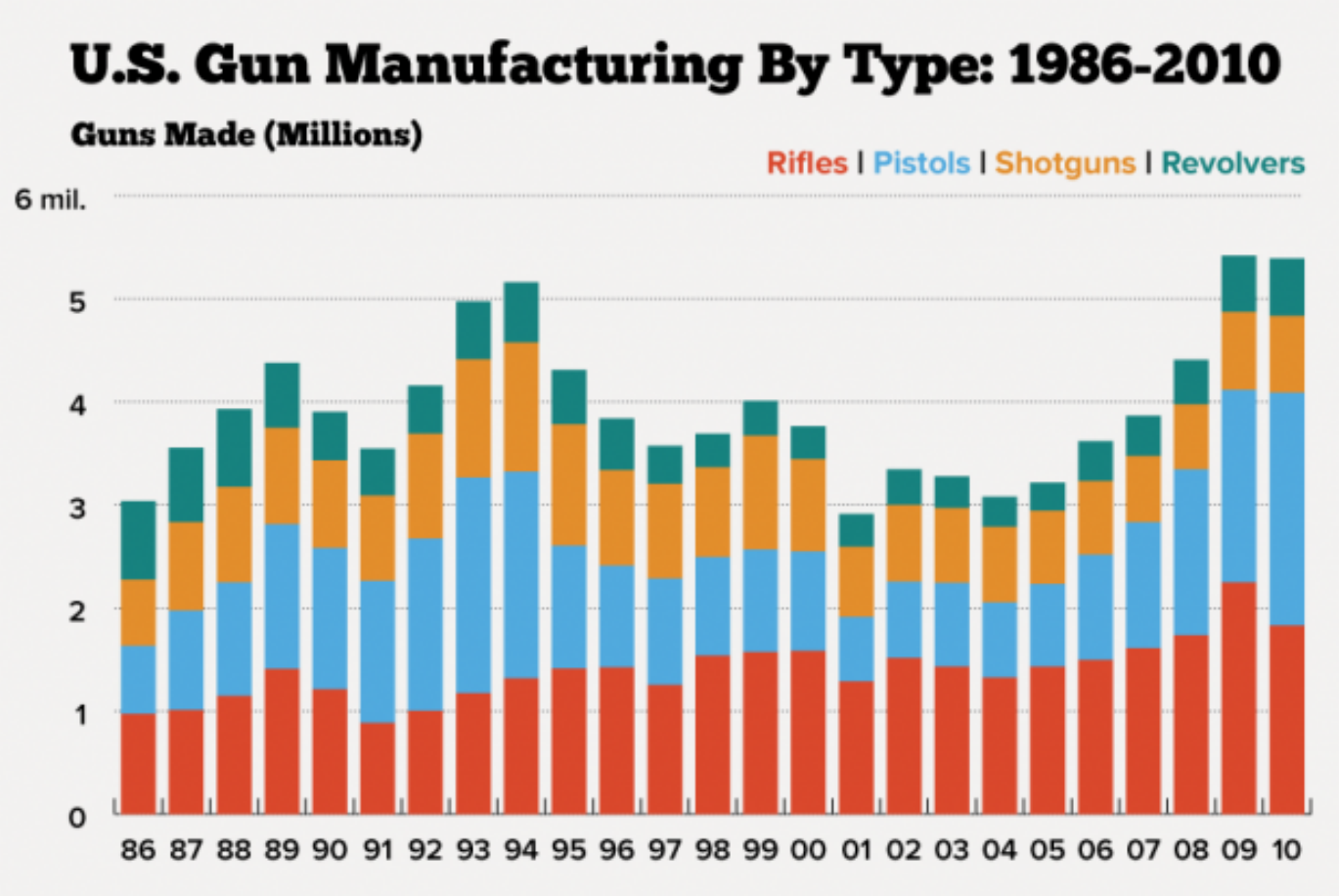
What we learn today

- What are the different ways to represent multiple categorical variables using bar charts?
- What types of tables are there and why do we use them?
- What are conditional statistics?
- Can we relate tables to their associated bar charts?

Bar Charts


Stacked
Clustered
Conditional
count
Region
Far West
Great Lakes
Mid East
New England
Plains
Rocky Mountains
South East
South West
Type
Private
Public

Stacked Bar Example



https://stackoverflow.com/questions/64267754/plotting-a-time-series-stacked-bar-chart

Descriptive Statistics – Categorical Variables

Univariate categorical variables are often presented in *tables*

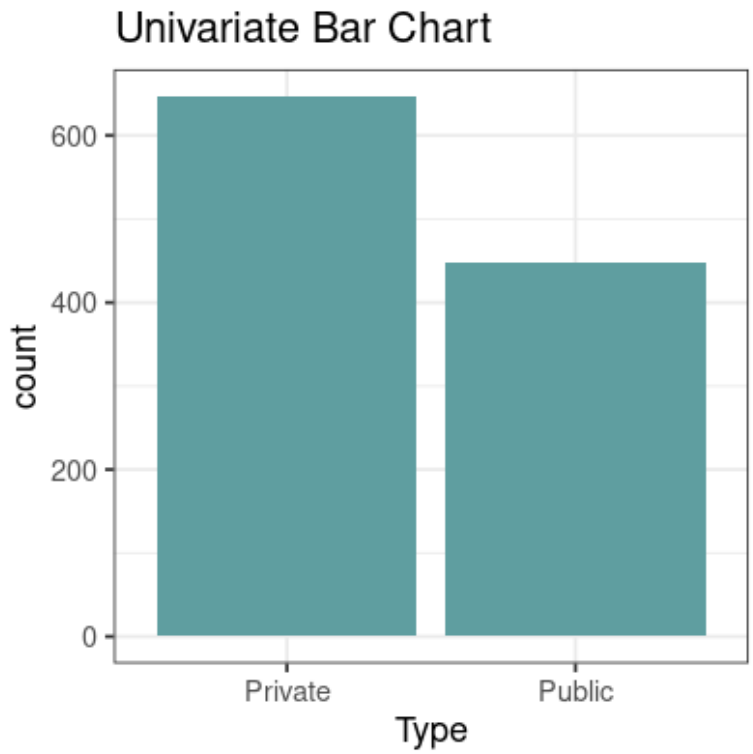
- **Frequencies:** counts how many of each case belongs to a particular category
- **Proportions:** fractions based upon frequencies, also called *relative frequencies*

Frequency table:

	Frequency
Private	647
Public	448

Table of proportions:

	Proportion
Private	0.591
Public	0.409



Bivariate Bar Charts

Just as we did when looking at graphical summaries, we tend to designate variables as being either *explanatory* or *response* variables

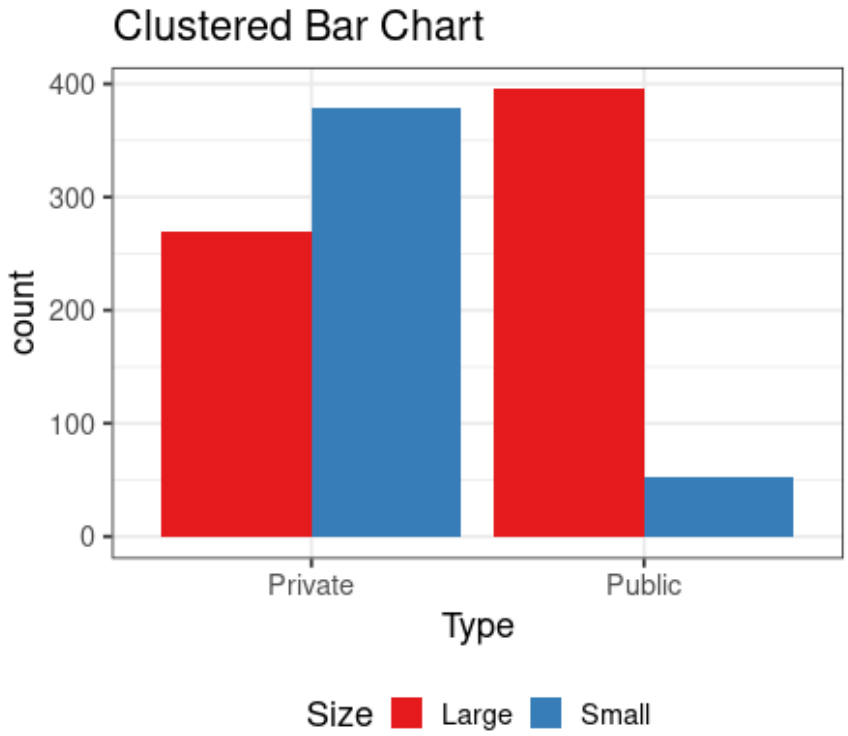
Again, this is **not** causal

We tend to think of these relationships *conditionally* when discussing categorical variables, a term we will return to shortly

Descriptive Statistics – Categorical Variables

Two-way frequency table:

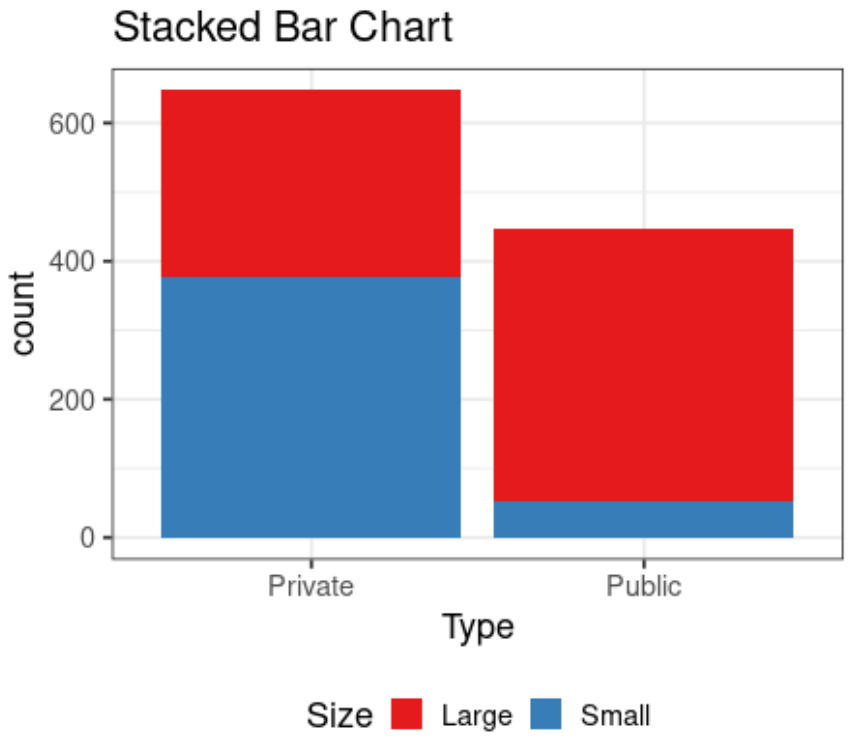
	Small	Large
Private	378	269
Public	53	395



Descriptive Statistics – Categorical Variables

Often these tables include margin sums as well

	Small	Large	Sum
Private	378	269	647
Public	53	395	448
Sum	431	664	1095



Descriptive Statistics – Categorical Variables

Two-way table of proportions

	Small	Large
Private	0.3452	0.2457
Public	0.0484	0.3607

"36% of all schools are large public schools"

Conditional Statistics

A **conditional statistic** is a statistic derived from one or more variables for all observations sharing a value of another variable

- "What is the relationship between admission rate and median ACT *given* that the school is private"
- "What is the predicted weight of an individual *given* that they are 6ft tall"
- "What is the proportion of public schools *given* that we are looking at the Plains region"

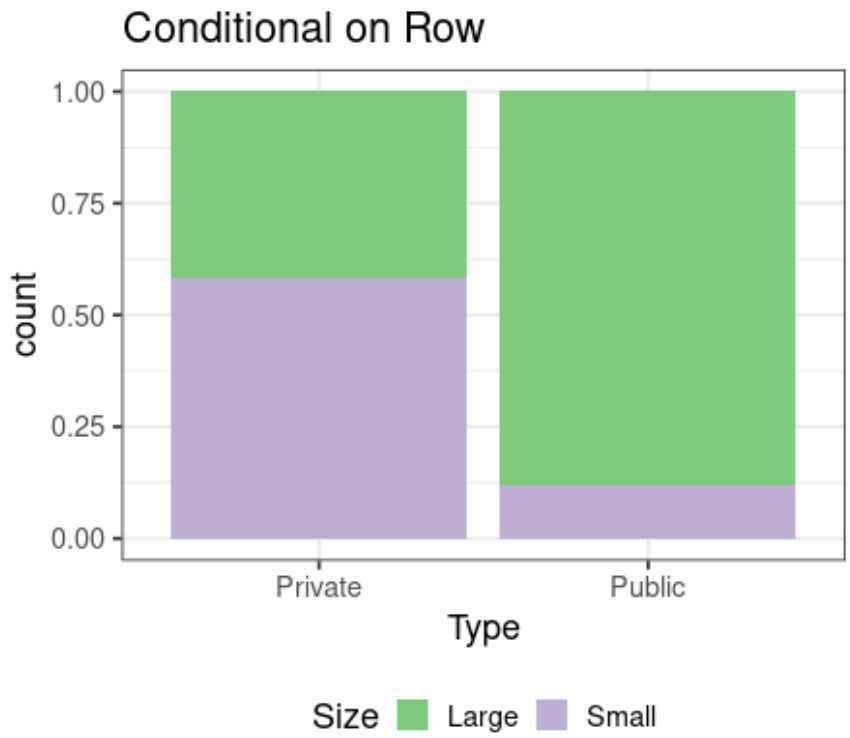
Note that we typically condition on the *explanatory* variable

Descriptive Statistics – Row Proportions

"88% of public schools are considered large"

"Given that a school is a public school, 88% of them are considered large"

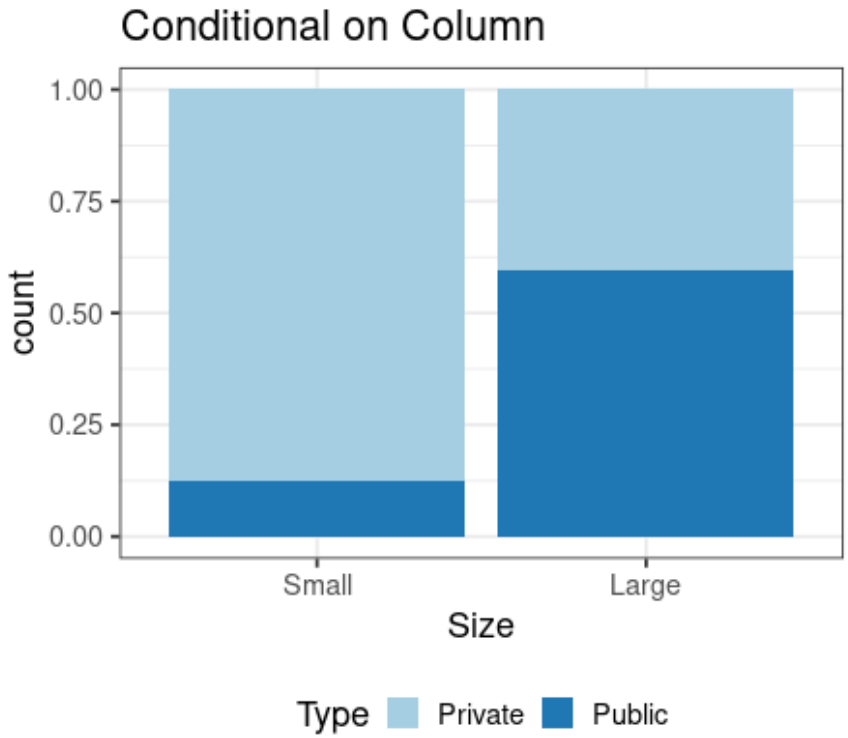
	Small	Large
Private	0.5842	0.4158
Public	0.1183	0.8817



Descriptive Statistics – Column Proportions

"12% of small colleges are public"

	Small	Large
Private	0.8770	0.4051
Public	0.1230	0.5949



Example

The two-way table below describes the survival of crew members and first class passengers aboard the Titanic

	Survived	Died
Crew	212	673
First Class	203	122

1. Given that an individual survived, is it more likely that they were a crew member or a passenger in first class?
2. Given that an individual was a crew member, is it more likely that they survived or died?
3. Which group was more likely to survive the shipwreck?

Summary

- Types of charts
  - Stacked
  - Clustered
  - Conditional
- Types of Tables
  - One and two-way tables
  - Frequency and proportions
  - Which associated with which plots?
- Association for categorical variables