Descriptive Statistics (Categorical)

Grinnell College

February 5, 2024

Review

What we learn today

- How do visualizations and descriptive statistics differ?
- 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?

Descriptive Statistics

Data visualizations – *qualitative* summary

- "X and Y have a weak positive linear relationship"

Descriptive statistics – *quantitative* summary

- "X and Y have a correlation coefficient of $r = 0.34$"

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

Descriptive Statistics – Categorical Variables

Bivariate categorical variables are often presented in a two-way table

Two-way frequency table:

Region	Private	Public
Far West	59	45
Great Lakes	125	64
Mid East	126	72
New England	44	27
Plains	84	42
Rocky Mountains	8	22
South East	163	130
South West	38	46

Descriptive Statistics – Categorical Variables

Often these tables include margin sums as well

	Private	Public	Total
Far West	59	45	104
Great Lakes	125	64	189
Mid East	126	72	198
New England	44	27	71
Plains	84	42	126
Rocky Mountains	8	22	30
South East	163	130	293
South West	38	46	84
Total	647	448	1095

Descriptive Statistics – Categorical Variables

Two-way table of proportions

Region	Private	Public
Far West	0.054	0.041
Great Lakes	0.114	0.058
Mid East	0.115	0.066
New England	0.040	0.025
Plains	0.077	0.038
Rocky Mountains	0.007	0.020
South East	0.149	0.119
South West	0.035	0.042

"2% of schools are public schools located in the Rocky Mountains"

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

"66% of schools in the Plains are private schools"

	Private	Public
Far West	0.567	0.433
Great Lakes	0.661	0.339
Mid East	0.636	0.364
New England	0.620	0.380
Plains	0.667	0.333
Rocky Mountains	0.267	0.733
South East	0.556	0.444
South West	0.452	0.548

Descriptive Statistics – Column Proportions

"13% of private schools are located in the Plains"

	Private	Public
Far West	0.091	0.100
Great Lakes	0.193	0.143
Mid East	0.195	0.161
New England	0.068	0.060
Plains	0.130	0.094
Rocky Mountains	0.012	0.049
South East	0.252	0.290
South West	0.059	0.103

Example

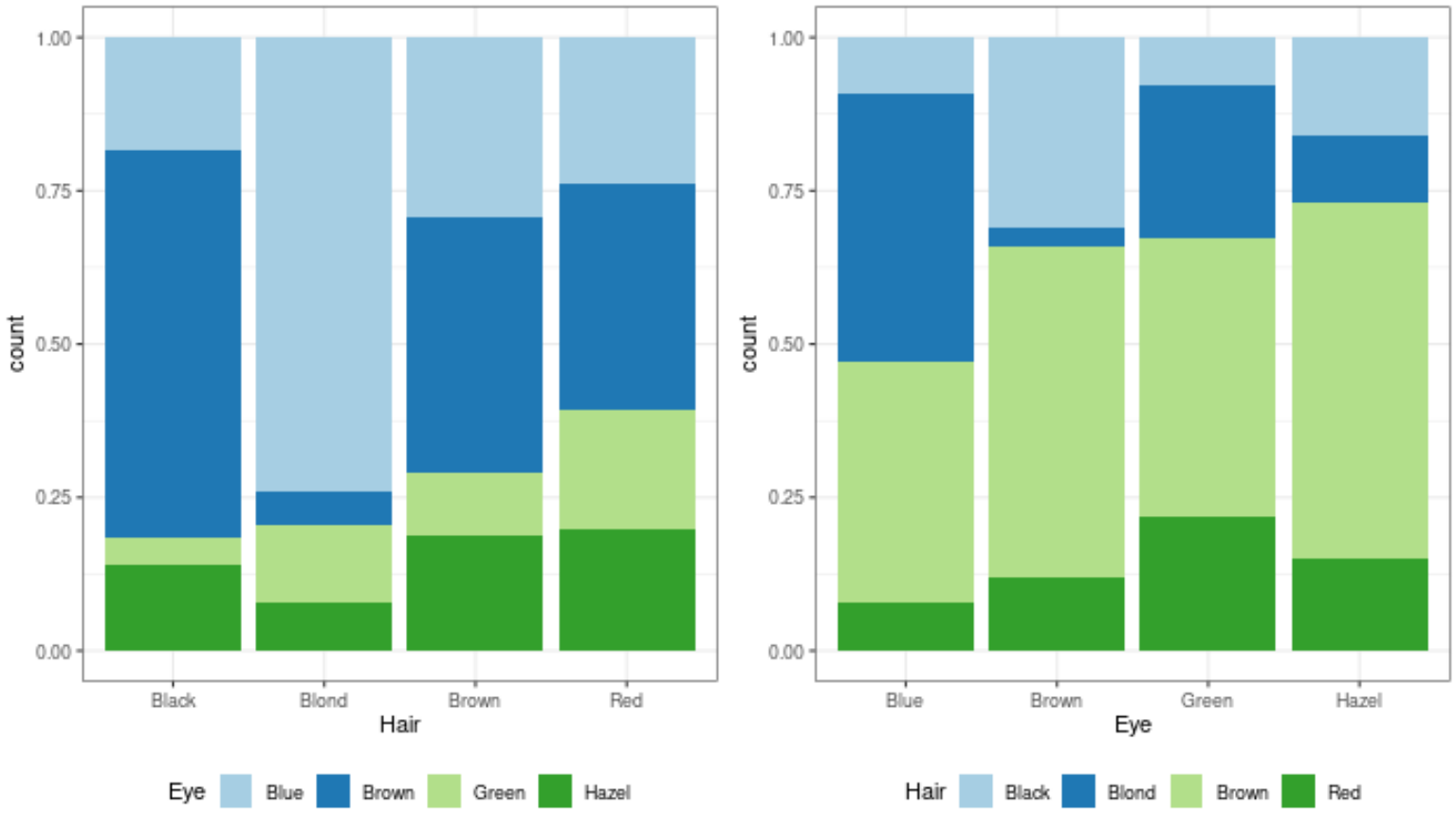
The two-way table below describes the survival of crew members and first class passengers aboard the Titanic

	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?

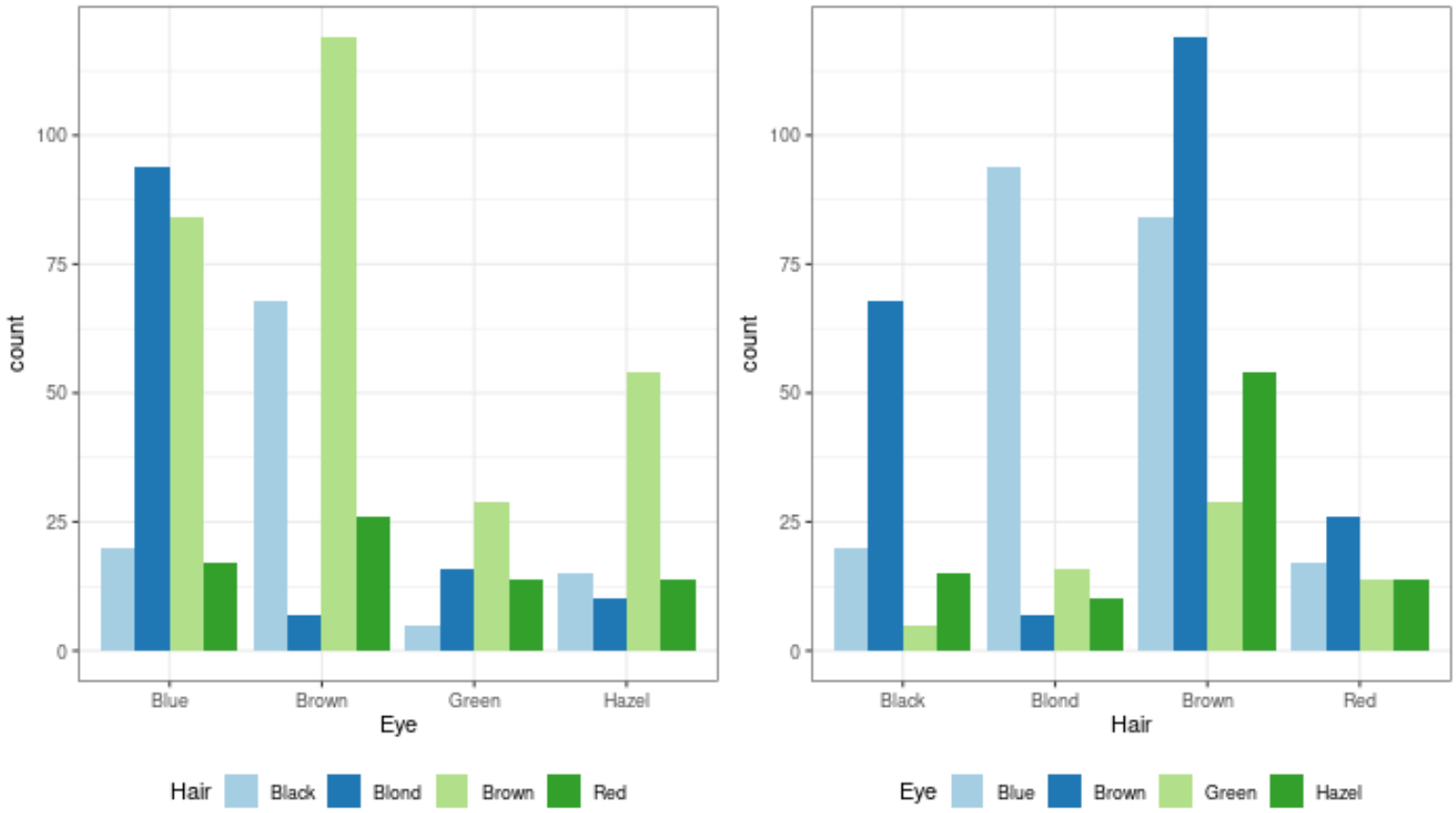
Example 2

	Blue	Brown	Green	Hazel
Black	20	68	5	15
Blond	94	7	16	10
Brown	84	119	29	54
Red	17	26	14	14



Example 2 Cont.

	Blue	Brown	Green	Hazel
Black	20	68	5	15
Blond	94	7	16	10
Brown	84	119	29	54
Red	17	26	14	14



Contingency Tables

A **contingency table** is a special two-way table in which both categorical variables have a binary response

	Event	Non-Event
Exposure	A	B
No Exposure	C	D

Odds

When dealing with a binary event, we often speak in terms of **odds**, a *ratio* of "number of successes" to "number of failures"

$$\#\text{ success} : \#\text{ failure}$$

This is distinct from the idea of **probabilities**, which give a ratio of the "number of successes" to the number of possible outcomes

$$\begin{aligned}\#\text{ success} &: \#\text{ total outcomes} \\ &: \#\text{ success} + \#\text{ failure}\end{aligned}$$

Odds

Suppose we have a 6-sided die, and we are interested in rolls that land on either 1 or 2 (success)

$$\text{Die} = \{1, 2, 3, 4, 5, 6\}$$

- The *probability* of rolling a 1 or 2 is $1/3$
  1. There are 6 possible outcomes
  2. There are 2 possible successes
  3. Probably is $2 \ / \ 6 = 1/3$
- The *odds* of rolling a 1 or 2 are 2:4 (or 1:2)
  1. There are 2 possible successes
  2. There are 4 possible failures
  3. The odds of success are 2:4 (or 1:2)

Odds Ratio

An **odds ratio** is the ratio of odds between two groups

	Event	Non-Event
Exposure	A	B
No Exposure	C	D

- The odds of an event for the exposure group are A:B (or A/B)
- The odds of an event for the no exposure gropu are C:D (or C/D)

The *odds* ratio for these groups is then the ratio of their odds:

$$OR = \frac{A:B}{C:D} = \frac{A/B}{C/D} = \frac{A \times D}{B \times C}$$

OR details

Discussion

Odds and Odds Ratio Example

A report published in 1988 summarizes results of a Harvard Medical School clinical trial determining effectiveness of asprin in preventing heart attacks in middle-aged male physicians

Treatment Status	Myocardial Infarction	
	Attack	No Attack
Placebo	189	10,845
Asprin	104	10,933

- Odds of having a heart attack for placebo:
- Odds ratio for treatment and infarction:
- Associated?

Review

- How do visualizations and descriptive statistics differ?
- 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?