

MATH 201: Lecture 3b Handout

Section 2.1 Considering categorical data

Name: _____ Date: _____

Learning goals for today

By the end of this lecture, you should be able to:

- Construct and interpret frequency and relative frequency tables.
- Create and interpret bar plots.
- Compare categorical distributions across groups.

Ways to summarize a single categorical variable

Fill in the blanks for each definition:

_____ is the number/count of observations in each category.

_____ is the proportion of observations in each category.

$$\text{Relative Frequency} = \frac{\text{Frequency}}{\text{Total Number of Observations}}$$

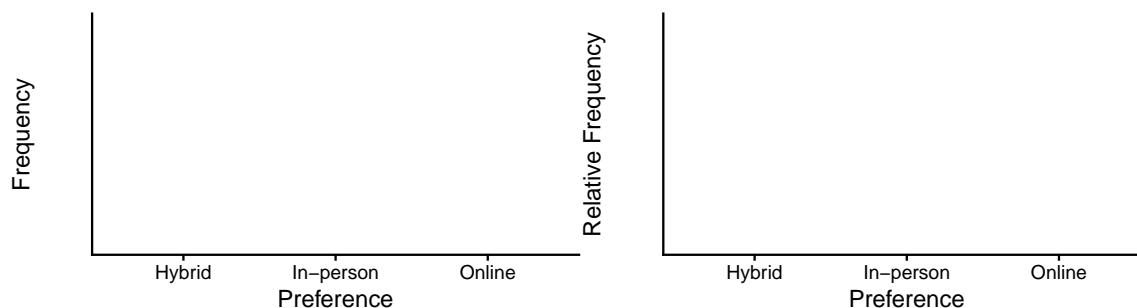
Ex: Say I surveyed 10 students asking if they preferred online, hybrid, or in-person class. Their responses were:

In-person, Hybrid, Hybrid, In-person, In-person, Online, In-person, Online, Online, In-person

Calculate the frequencies and relative frequencies:

Bar plots/charts display frequencies (or relative frequencies) for categories.

Draw a bar plot for the class preference data.



Summarizing two categorical variables

	Hybrid	In-person	Online
Less than 2 years	18	22	10
2+ years	12	28	20

Suppose we surveyed more people and also recorded how long they have attended CSUCI. The **two-way contingency table** above shows this data.

1. Add row and column totals to the table.
2. How many students have attended CSUCI for less than 2 years?
3. How many students have attended CSUCI for less than 2 years and prefer hybrid class?
4. Among students who have attended CSUCI for less than 2 years, what proportion prefer:

$$\text{hybrid} = \frac{18}{50}$$

$$\text{in-person} =$$

$$\text{online} =$$

5. Among students who have attended CSUCI for more 2+ years, what proportion prefer:

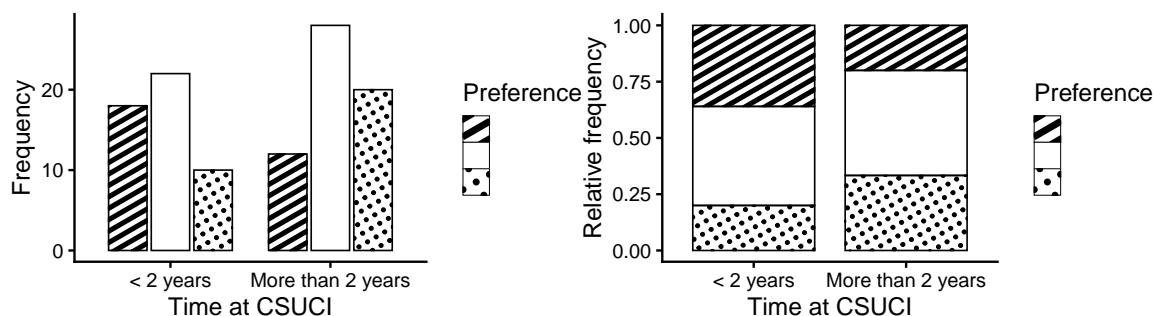
$$\text{hybrid} = \frac{12}{60}$$

$$\text{in-person} =$$

$$\text{online} =$$

6. Do the proportions for questions 4 and 5 look similar across the two groups?

Dodged bar plots display frequencies and **standardized bar plots** display relative frequencies for two categorical variables.



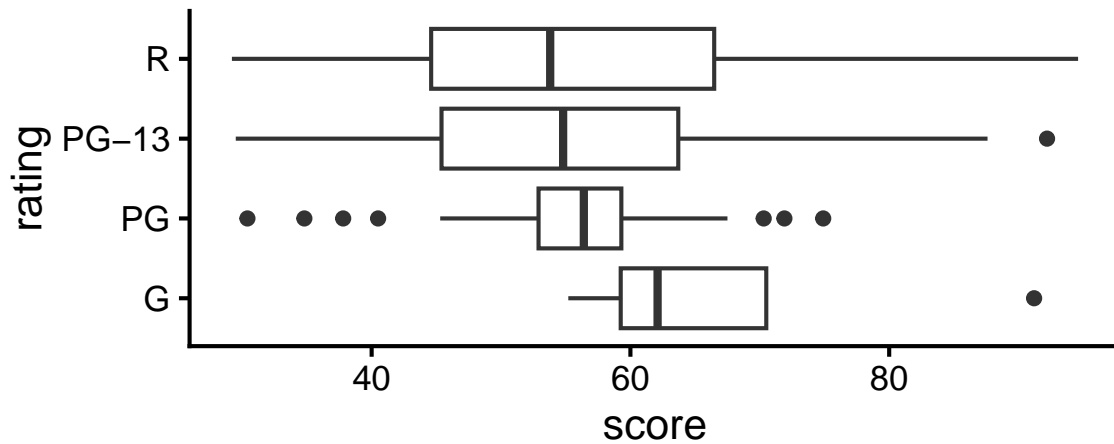
Fill in the missing legend labels.

Recall that we say two variables are **associated** if knowing one variable gives some information about the other. Would you say class preference appears to be associated with time at CSUCI? Why or why not?

Examining one numerical and one categorical variable

Side-by-side box plots are often helpful for examining the relationship between a numerical and a categorical variable.

In this example we will use a dataset of movies. For each movie we have information on their critic score (from 0-100), rating (G, PG, PG-13, or R), and box office earnings (in millions of dollars).

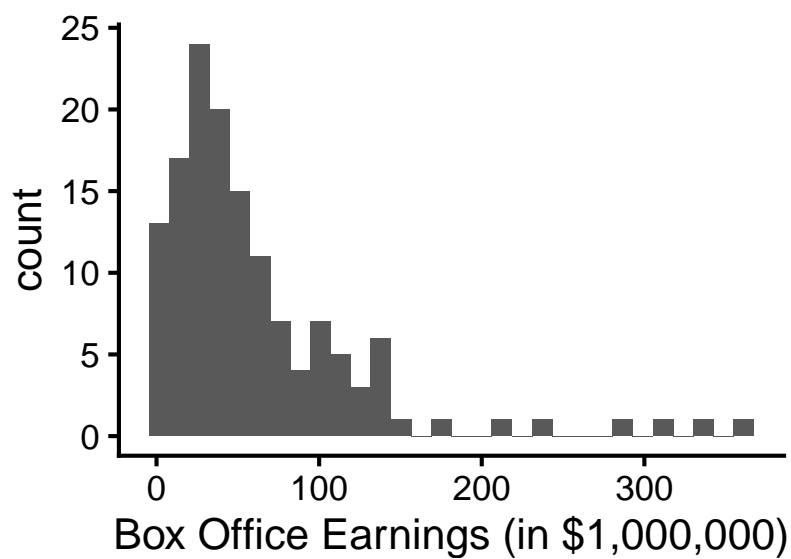


List 3 factual observations about the plot shown above.

- 1.
- 2.
- 3.

Which ratings categories appear to be similar in critic score? Which appear to be the least similar? Does this make sense?

More practice



How would you describe the distribution of box office earnings?

If we doubled the dataset by duplicating every observation, how would the side-by-side box plot of ratings versus critic score change, if at all?

If we doubled the dataset by duplicating every observation, how would the histogram of box office earnings change, if at all?