STAT 22000 Lecture Slides
Exploring Categorical Data

Yibi Huang
Department of Statistics
University of Chicago

Outline

This set of slides cover Section 1.7 in the text.

- Ways to summarize of a single categorical variable
    - Frequency tables
    - Barplots, pie charts
- Ways to summarize of relationships between two categorical variables
    - two-way contingency tables
    - segmented barplots, standardized segmented barplots, mosaic plot

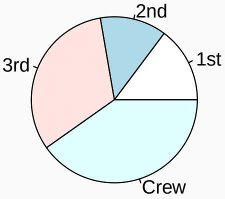
Bar Graphs and Pie Charts

Graphs for Categorical Variables

A categorical variable is summarized by a table showing the *count* or the *percentage* of cases in each category, and is often displayed by a *bar plot* or a *pie chart*.

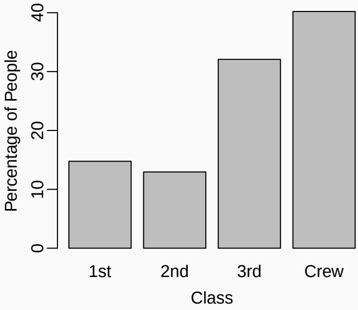
Ex: Passengers on Titanic

Class	Freq	Percent
1st	325	14.8%
2nd	285	12.9%
3rd	706	32.1%
Crew	885	40.2%
Total	2201	100%



Bar plots

A *bar plot* is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a *relative frequency bar plot*.

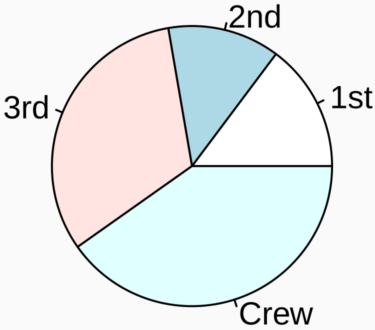


How are Bar Plots Different From Histograms?

- Bar plots are used for displaying distributions of categorical variables, while histograms are used for numerical variables.
- The horizontal axis in a histogram is a number line, hence the order of the bars cannot be changed, while in a bar plot the categories can be listed in any order (though some orderings make more sense than others, especially for ordinal variables.)

Why We Recommend Bar Plots Over Pie Charts?

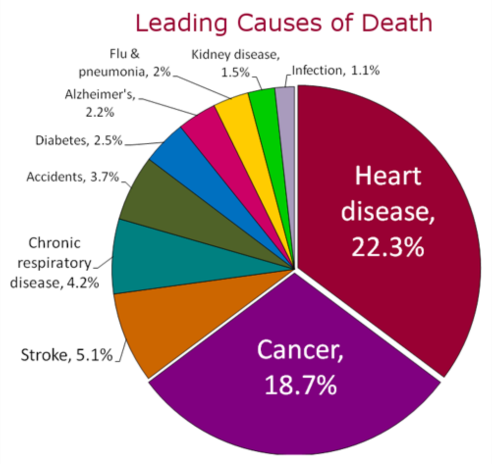
In a pie chart, the **areas** of slices represents the **percentages** of categories. However, it is generally more difficult to compare group sizes in a pie chart than in a bar plot, especially when categories have nearly identical counts or proportions



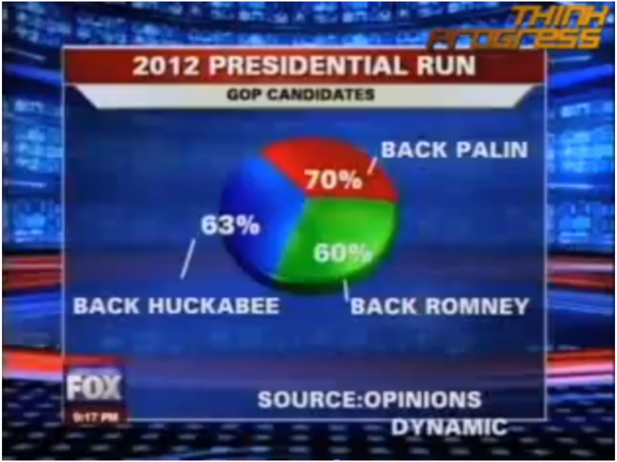
Without looking at the counts, can you tell which class have fewest people from the pie?

Why We Recommend Bar Plots Over Pie Charts?

It's much easier to make a wrong pie chart than a wrong bar plot.
In a pie chart, the categories must make up a **whole**. There is no such restriction for a bar plot.



Another Wrong Pie Chart



http:// www.youtube.com/ watch?v=-rbyhj8uTT8

Two-Way Contingency Tables

Two-Way Contingency Tables

A table that summarizes data for two categorical variables is called a *contingency table*.

E.g., breakdown of people on Titanic by class and survival status

		Died	Survived	Total
Class	1st	122	203	325
	2nd	167	118	285
	3rd	528	178	706
	Crew	673	212	885
	Sum	1490	711	2201

The marginal totals give the distributions of the two variables, e.g.,

- overall, 1490 died and 711 survived
- there were 325, 285, and 706 passengers in the 1st, 2nd and 3rd classes, and 885 crew members

Overall Proportions

Dividing the cell counts in a contingency table by the overall total, we get the proportions of observations in the combinations of the two variables.

		Survived		
		No	Yes	Total
Class	1st	$122/2201 \approx 0.06$	$203/2201 \approx 0.09$	$325/2201 \approx 0.15$
	2nd	$167/2201 \approx 0.08$	$118/2201 \approx 0.05$	$285/2201 \approx 0.13$
	3rd	$528/2201 \approx 0.24$	$178/2201 \approx 0.08$	$706/2201 \approx 0.32$
	Crew	$673/2201 \approx 0.31$	$212/2201 \approx 0.10$	$885/2201 \approx 0.40$
	Sum	$1490/2201 \approx 0.68$	$711/2201 \approx 0.32$	1

e.g., of people on Titanic

- $122/2201 \approx 6\%$ were in the 1st class and died in the disaster
- $212/2201 \approx 10\%$ were survived crew members

Note the marginal totals give the distributions of the two variables, e.g.,

- Overall, $711/2201 \approx 32\%$ of the people survived

Row Proportions

The row proportions (cell counts divided by the corresponding row totals) give the proportion of people survived in the four classes.

		Survived		
		No	Yes	Total
Class	1st	$122/325 \approx 0.38$	$203/325 \approx 0.62$	1
	2nd	$167/285 \approx 0.59$	$118/285 \approx 0.41$	1
	3rd	$528/706 \approx 0.75$	$178/706 \approx 0.25$	1
	Crew	$673/885 \approx 0.76$	$212/885 \approx 0.24$	1

e.g.,

- $203/325 \approx 62\%$ of people in the 1st class survived.
- $178/706 \approx 25\%$ of people in the 3rd class survived.

Column Proportions

The column proportions (dividing cell counts by the corresponding column totals) give the proportion of people survived in each of the four classes.

		Survived	
		No	Yes
Class	1st	$122/1490 \approx 0.08$	$203/711 \approx 0.29$
	2nd	$167/1490 \approx 0.11$	$118/711 \approx 0.17$
	3rd	$528/1490 \approx 0.35$	$178/711 \approx 0.25$
	Crew	$673/1490 \approx 0.45$	$212/711 \approx 0.30$
	Sum	1	1

- Among those who survived, $203/711 \approx 29\%$ were in the 1st class.
- Among those who died, $673/1490 \approx 45\%$ were crew members

Independence of Two Categorical Variables

If the row proportions do not change from row to row, we say the two categorical variables are *independent*. Otherwise, we say they are *associated*.

E.g., if the survival rates do not change from class to class, we say 'survival' is independent of 'class'. In the Titanic data, the survival of passengers is associated with the class they were in because the survival rates differ substantially from class to class.

We can also define two categorical variables to be independent if the column proportions do not vary from column to column since the two conditions are equivalent (why?)

Exercise

The table below shows the breakdown of cases of injuries in the U.S in a certain year. by circumstance and gender[1]. Counts are in millions.

	Circumstance			
Gender	Work	Home	Other	Total
Male	8.0	9.8	17.8	35.6
Female	1.3	11.6	12.9	25.8
Total	9.3	21.4	30.7	61.4

- What proportion of injury cases occurred at work?
  $9.3/61.4 \approx 0.15$
- What proportion of injury cases occurred at work and on women? $1.3/61.4 \approx 0.02$

[1] Source: *Vital and Health Statistics* published by the *National Center for Health Statistics*

Practise (Cont'd)

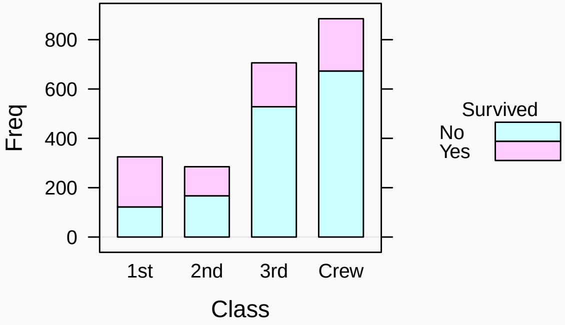
	Circumstance			
Gender	Work	Home	Other	Total
Male	8.0	9.8	17.8	35.6
Female	1.3	11.6	12.9	25.8
Total	9.3	21.4	30.7	61.4

- Among all injury cases occurred on women, what proportion occurred at work? $1.3/25.8 \approx 0.05$
- Among all injury cases occurred at work, what proportion occurred on women? $1.3/9.3 \approx 0.14$
- Is the circumstance of injury cases independent of the gender of the victims? No, only 5% of injury cases on women occurred at work, compared with $8.0/36.5 \approx 22\%$ of cases on men occurred at work.

Segmented Bar and Mosaic Plots

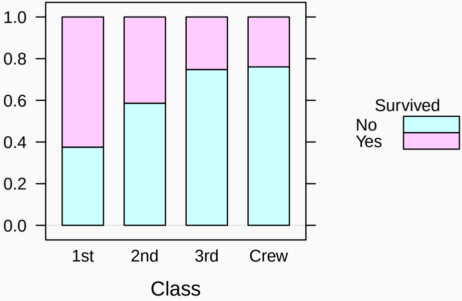
Segmented Bar Plots

	Survived		
Class	No	Yes	Total
1st	122	203	325
2nd	167	118	285
3rd	528	178	706
Crew	673	212	885
Sum	1490	711	2201



Standardized Segmented Bar Plots

		Survived		
		No	Yes	Total
Class	1st	0.38	0.62	1
	2nd	0.59	0.41	1
	3rd	0.75	0.25	1
	Crew	0.76	0.24	1



Standardized segmented bar plots are convenient for comparing row proportions, and determining whether the two variables are independent.

However, the information of row totals is lost after standardization.

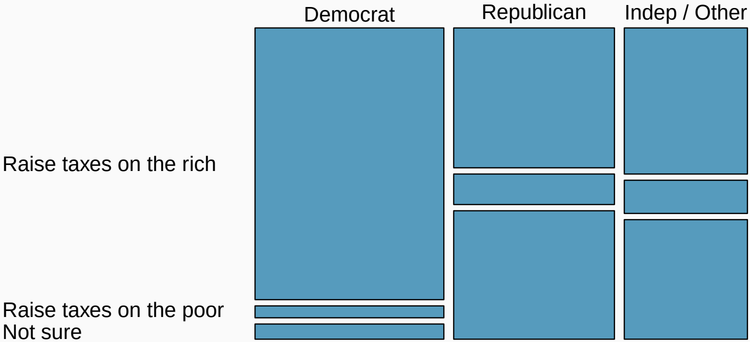
Mosaic Plots

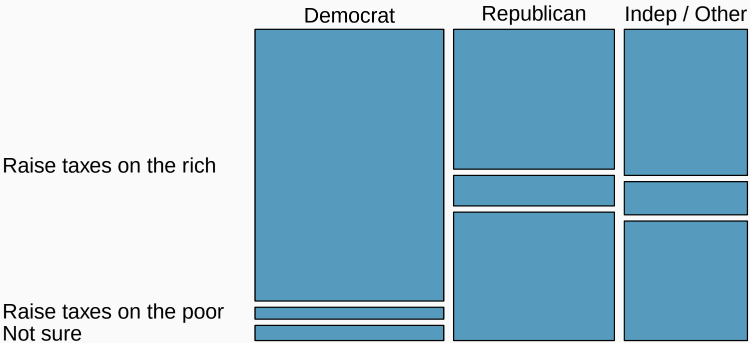
- bar widths = row totals
- segment lengths within a bar = row proportions

$$\boxed{\text{segment area}} = (\text{barwidth}) \times (\text{segment length})$$
$$= \text{row total} \times (\text{row proportion})$$
$$= \text{row total} \times \frac{\text{cell count}}{\text{row total}} = \boxed{\text{cell count}}$$

Exercise 1.68 Raise Taxes on the Rich or the Poor

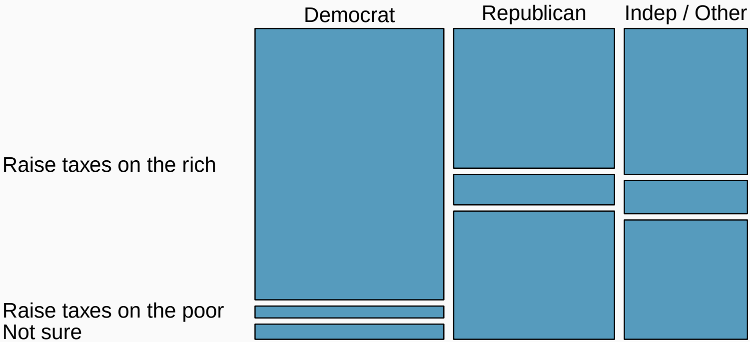
The mosaic plot below shows the relationship between political party affiliation and views on whether it's better to raise taxes on the rich or on the poor for a random sample of registered voters taken nationally in 2015.





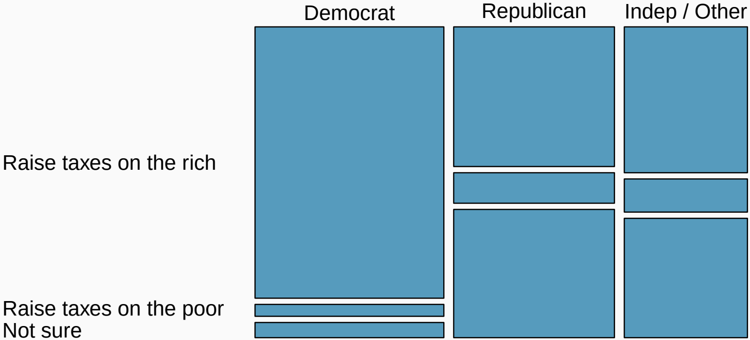
Which political party identification is least common in the sample, Democrats, Republicans, or Indep/Other?

Ans: Indep/Other.



Based on this sample, which political party identification had the highest percentage supported raising taxes on the rich? Which had the lowest?

Ans: Democrats the highest, Republicans the lowest.



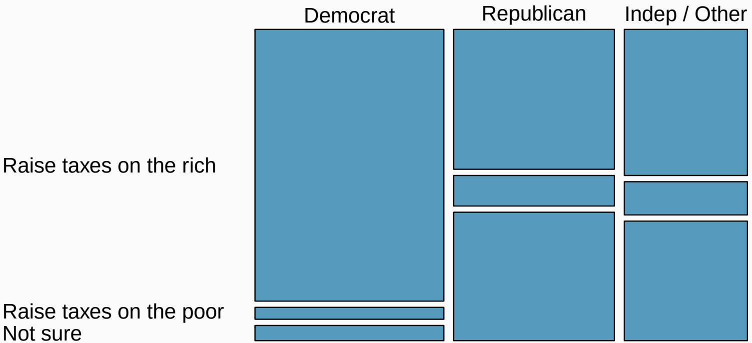
What percentage of Democrats (in this sample) supported raising taxes on the rich?

(a) below 25%

(b) between 25% and 50%

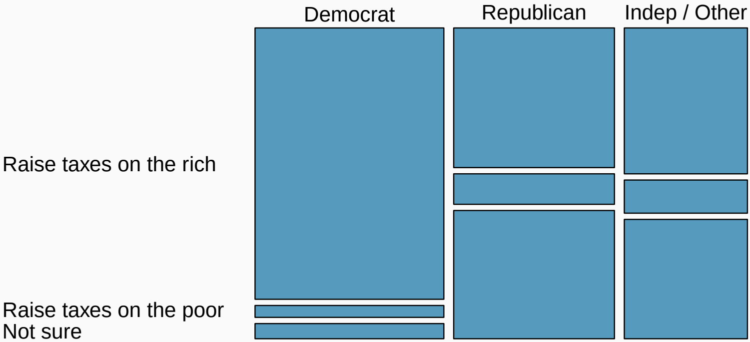
(c) between 50% and 75%

(d) over 75%



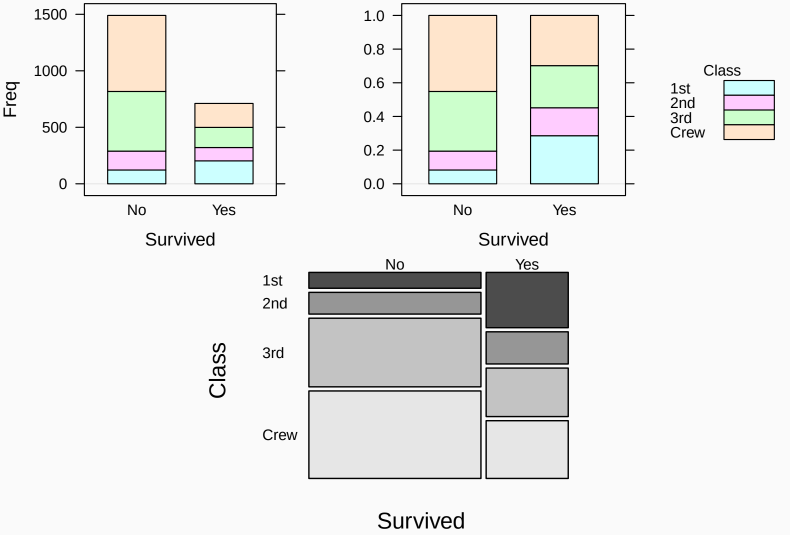
In this sample, which of the following groups contains the greatest number of subjects?

(a) Democrats who supported raising taxes on the rich.
(b) Democrats who supported raising taxes on the poor.
(c) Republicans who supported raising taxes on the rich.
(d) Republicans who supported raising taxes on the poor.



Based on the mosaic plot, do views on raising taxes and political affiliation appear to be independent?

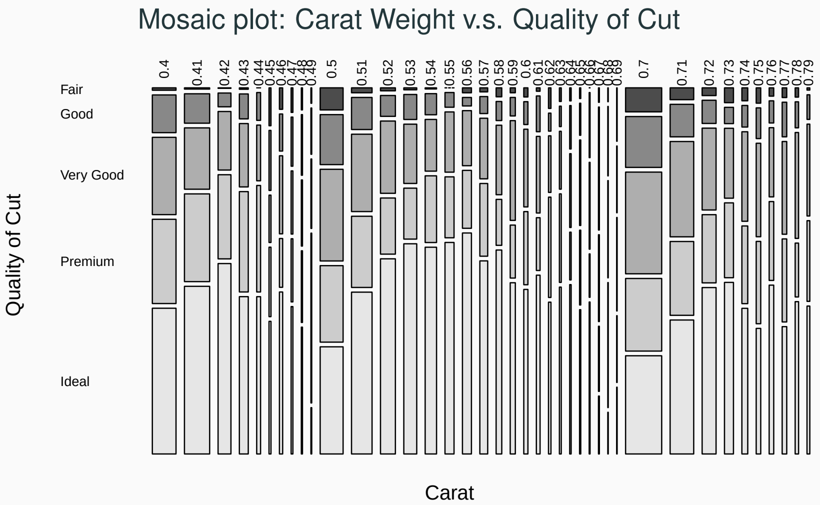
Instead of looking at survival rates in the four classes, we can also look at the breakdown of the four classes among those who survived and among those who died.



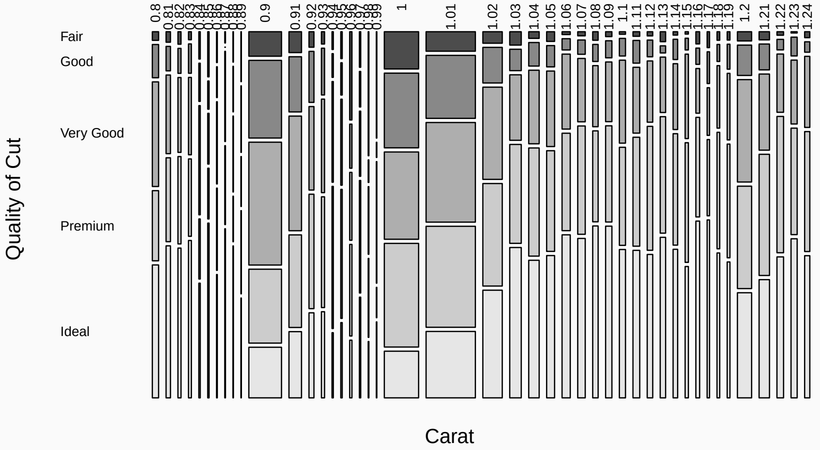
Ways to Inspect Relationships Between Variables

- numerical v.s. numerical
  - scatterplots
- categorical v.s. categorical
  - contingency tables
  - segmented barplots, standardized segmented barplots, mosaic plot
- categorical v.s. numerical
  - side-by-side boxplots
  - histograms by group on the same horizontal axis

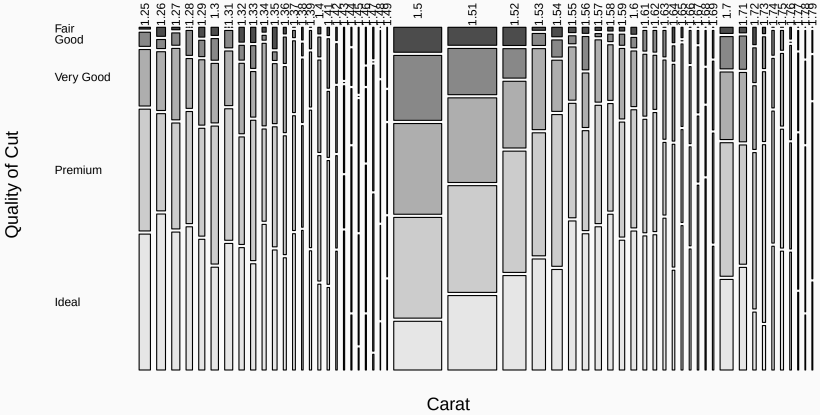
Example (Diamonds)



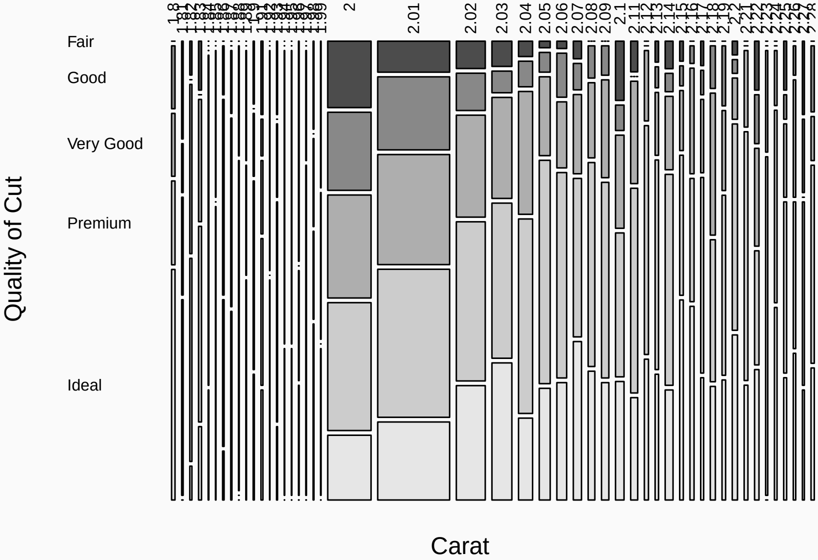
Example (Diamonds)



Example (Diamonds)



Example (Diamonds)



Example (Diamonds)

From the mosaic plots, we can see the proportion of low-quality cut diamonds increases substantially whenever the carat weight of diamonds reaches those benchmarks (0.5, 0.7, 0.9, 1, 1.2, 1.5, 2,. . .). Diamonds with carat weights right above those benchmarks generally have better quality of cut then those just at those benchmarks.

Possible reasons:

Diamond cutters would want to get the heaviest diamond out of a rough stone whenever possible. They might increase the depth of diamonds to increase the carat weight, but result in a loss of brilliance due to light leakage.