SUMMARIZING CATEGORICAL DATA

- ➢ Frequency Tables
- ➢ Pie Charts and Bar Graphs
- ➢ Cross-Classification Tables
- ➢ Side-by-Side and Segmented Bar Graphs
- ➢ Mosaic Plots
- ➢ Simpson's Paradox

University of Pittsburgh

SUMMARIZING A SINGLE CATEGORICAL VARIABLE

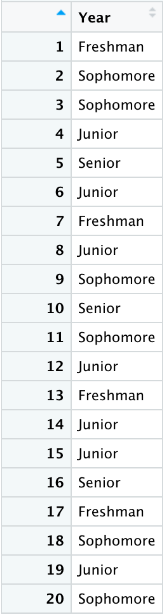
- **Count:** number of observations in a category
- **Proportion:** count in a category divided by total number of observations
- **Percentage:** proportion as decimal times 100%

- **Frequency table:** table of counts for each category
  - Sum to total number observations if categories do not overlap
- **Relative frequency table:** table of proportions or percentages for each category
  - Sum to 1 for proportions and 100% for percentages if categories do not overlap

EXAMPLE: SUMMARIZING A SINGLE CATEGORICAL VARIABLE

- **Scenario:** Random sample of 20 students taking STAT 1000 were asked what year in school they are in
- **Task:** Complete the frequency and relative frequency tables.

Year	Count	Proportion
Freshman		
Sophomore		
Junior		
Senior		
Total		



	Year
1	Freshman
2	Sophomore
3	Sophomore
4	Junior
5	Senior
6	Junior
7	Freshman
8	Junior
9	Sophomore
10	Senior
11	Sophomore
12	Junior
13	Freshman
14	Junior
15	Junior
16	Senior
17	Freshman
18	Sophomore
19	Junior
20	Sophomore

SUMMARIZING A SINGLE CATEGORICAL VARIABLE

- Parameters and statistics are denoted by different notations
  - Each parameter and statistic is given a different symbol

- For proportions:
  - p: Population proportion (parameter) that describes the proportion of observations in a category in an entire population
    - Typically unknown
  - $\hat{p}$: Sample proportion (statistic) that describes the proportion of observations in a category a sample
    - Representative sample estimates the population proportion p quite well

EXAMPLE: USING CATEGORICAL DATA

- **Scenario:** Survey of 796 college students found 288 reported binge drinking at some point in the past month
- **Question:** Why is this variable categorical?
- **Answer:** ________________ responses that are _______________
  - ______________ or _____

- **Question:** How can this data be summarized numerically?
- **Answer:** _____________ (or ____________)
  - ____________________ reported binge drinking

EXAMPLE: PARAMETER VS. STATISTIC

- **Scenario:** Survey of 796 college students found 288 reported binge drinking at some point in the past month
- **Question:** How should the proportion .362 be denoted?
- **Answer:** ___________ proportion: _________
  - ___________ describing a _________

- **Question:** How should the overall proportion of all college students who binge drink be denoted?
- **Answer:** ___________ proportion: ___
  - ___________ whose actual value is ________ → Cannot ______________

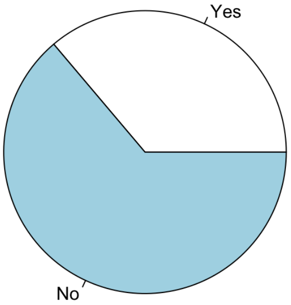
DISPLAYING A SINGLE CATEGORICAL VARIABLE

- **Pie chart:** shows proportion of observations in each category of a categorical variable where the size of the slice corresponds to the proportion
  - Should be used when:
    - Responses cannot __________ into more than one category
    - Want to show how the ________ is divided into categories
- **Bar graph:** displays counts (or proportions) for each category where the height of the bar corresponds to the count (or proportion) for the category
  - Can be used any time a pie chart can be used
  - Should be used when:
    - Responses can be provided for _________________ category (i.e. there is overlap)
    - __________________ are included in the comparison

EXAMPLE: DISPLAYING CATEGORICAL DATA

- **Scenario:** Survey of 796 college students found 288 reported binge drinking at some point in the past month
- **Question:** Why is a pie chart appropriate to display this data?
- **Answer:** Each student falls into _____________________________________

- **Question:** Can we conclude that more than one-third (33.3%) of all college students binge drink?
- **Answer:** ______
  - Indication this may be _________
  - _______ is only _____________________ 0.333
  - Only have a ___________ from ________________________

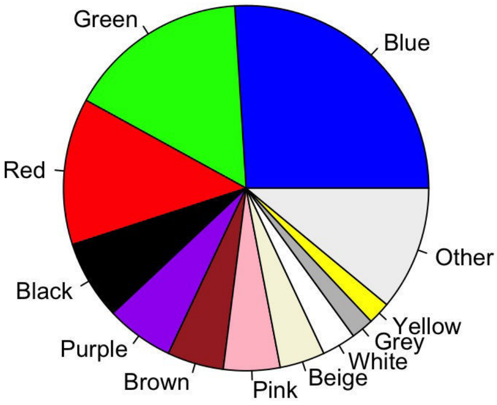


Looking Ahead: This would require ____________________.

EXAMPLE: PROBLEMS WITH PIE CHART

- **Scenario:** Random sample of adults asked for favorite color
- **Question:** What are the problems with using a pie chart?
- **Answer:** Difficult to…
  - ___________________________ categories
  - See categories with a ______________ of the chart

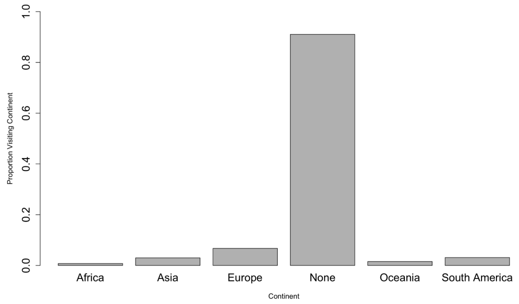
- **Takeaway:** Pie charts typically work well for viewing variables with up to ________ __________________.



EXAMPLE: BAR GRAPH

- **Scenario:** Survey of 10,000 United States residents asked which continents they have visited outside of North America

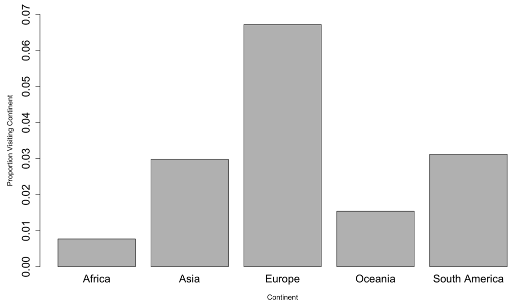
Continent	Count	Proportion
Europe	672	0.0672
South America	312	0.0312
Asia	298	0.0298
Oceania	154	0.0154
Africa	77	0.0077
None	9105	0.9105



- **Question:** Why is a bar graph more appropriate than a pie chart?
- **Answer:** People can visit ______________________________
  - Categories _________ so proportions sum to _____________

EXAMPLE: BAR GRAPH

- **Scenario:** Survey of 10,000 United States residents asked which continents they have visited outside of North America
- **Question:** Why is removing the bar for "None" appropriate?
- **Answers:** Other five continents are ________________________
  - Leaving it in _________ the graph

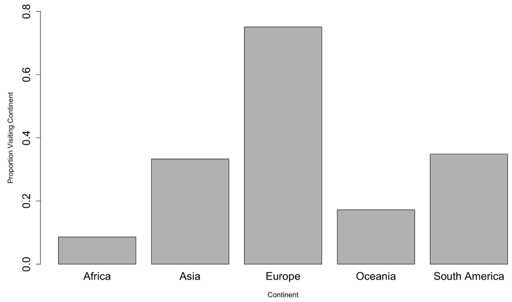


- **Question:** Are there any trends in visitation?
- **Answer:** Continents that are _________________ and have better _________________ tend to have _________ visitation

EXAMPLE: CONDITIONAL PROPORTIONS

- **Scenario:** Remove the people who have not left North America.

Continent	Count	Proportion
Europe	672	0.7508
South America	312	0.3486
Asia	298	0.3330
Oceania	154	0.1721
Africa	77	0.0860
Total	**895**	**1.00**



- **Question:** What does the 0.7508 mean for Europe?
- **Answer:** Of the ____ surveyed people who have traveled outside of North America, _________ of them have ___________________
  - Compared to _______ of __________ people who were surveyed.

MOTIVATION: COMPARING CATEGORICAL VARIABLES

- **Scenario:** Random sample of 424 people who recently joined a gym were asked for the reason they joined (fitness or weight loss) and their sex
- **Question:** What is the goal of this survey?
- **Answer:** Understand if there is a __________ between ____ and why a person __________________
- **Question:** What are the explanatory and response variables?
- **Answer:**
  - **Response:** ______________________
  - **Explanatory:** ___________

CROSS-CLASSIFICATION TABLE

- **Cross-classification table:** a display of counts, proportions, or percentages that compare two categorical variables
  - Each **cell** contains the value corresponding to the combination of values where the row and column intersect.
  - If analyzing a relationship:
    - Explanatory variable typically goes along rows
    - Response variable typically goes along columns
  - Also called **contingency tables** and **two-way tables**

EXAMPLE: CROSS-CLASSIFICATION TABLE

- **Scenario:** Random sample of 424 people who recently joined a gym were asked for the reason they joined and their sex
- **Question:** How many females joined for fitness reasons?
- **Answer:** ____
  - Intersection of the _____ for females and ________ for fitness

	Fitness	**Weight Loss**	**Total**
Male	98	114	**212**
Female	95	117	**212**
Total	**193**	**231**	**424**

- **Question:** Does there appear to be a relationship between sex and reason for joining a gym?
- **Answer:** _______
  - Males and females _________________ to join a gym for each reason
  - ____________ of each sex were surveyed and counts are _________

Motivation: Conditional Distribution

- **Scenario:** Compare year in school against if a student has a job

	Job	No Job	Total
Freshman	22	24	**46**
Sophomore	170	142	**312**
Junior	97	37	**134**
Senior	37	16	**53**
Total	**326**	**219**	**565**

- **Question:** Are sophomores the most likely to have a job?
- **Answer:** _______
  - Many __________________ were surveyed → Counts are ______________
  - Need to analyze __________________________

Conditional Distribution

- **Conditional distribution:** shows the proportion/percentage of observations in each response category given that the observation falls into a specific explanatory group
  - Divide count for cell inside table by **row sum** for explanatory category

Cross-Classification Table

	Job	No Job	Total
Freshman	22	24	**46**
Sophomore	170	142	**312**
Junior	97	37	**134**
Senior	37	16	**53**
Total	**326**	**219**	**565**

Conditional Distribution

	Job	No Job	Total
Freshman	47.83%	52.17%	**100.00%**
Sophomore	54.49%	45.51%	**100.00%**
Junior	72.39%	27.61%	**100.00%**
Senior	69.81%	30.19%	**100.00%**
Total	**57.70%**	**42.30%**	**100.00%**

Proportion of freshmen who have a job: $\frac{22}{46} = .4783$

Example: Cross-Classification Table

- **Scenario:** Compare year in school against if a student has a job

	Job	No Job	Total
Freshman	22	24	**46**
Sophomore	170	142	**312**
Junior	97	37	**134**
Senior	37	16	**53**
Total	**326**	**219**	**565**

	Job	No Job	Total
Freshman	47.83%	52.17%	**100.00%**
Sophomore	54.49%	45.51%	**100.00%**
Junior	72.39%	27.61%	**100.00%**
Senior	69.81%	30.19%	**100.00%**
Total	**57.70%**	**42.30%**	**100.00%**

- **Question:** Is there a relationship between year and having a job?
- **Answer:** ______
  - ________________ are more likely to have a job
  - __________________ in conditional percentages (____ vs. ____)

EXAMPLE: CROSS-CLASSIFICATION TABLE

- **Scenario:** Compare year in school against if a student has a job

	Job	No Job	Total
Freshman	22	24	**46**
Sophomore	170	142	**312**
Junior	97	37	**134**
Senior	37	16	**53**
Total	**326**	**219**	**565**

	Job	No Job	Total
Freshman	47.83%	52.17%	**100.00%**
Sophomore	54.49%	45.51%	**100.00%**
Junior	72.39%	27.61%	**100.00%**
Senior	69.81%	30.19%	**100.00%**
Total	**57.70%**	**42.30%**	**100.00%**

- **Question:** Are juniors more likely to have a job than seniors in general?
- **Answer:** ________________
  - Only ____________ sampled and ___________is only _____
  - __________ to make a ___________________ based on a sample

EXAMPLE: CROSS-CLASSIFICATION TABLE

- **Scenario:** Compare year in school against if a student has a job

	Job	No Job	Total
Freshman	22	24	**46**
Sophomore	170	142	**312**
Junior	97	37	**134**
Senior	37	16	**53**
Total	**326**	**219**	**565**

	Job	No Job	Total
Freshman	47.83%	52.17%	**100.00%**
Sophomore	54.49%	45.51%	**100.00%**
Junior	72.39%	27.61%	**100.00%**
Senior	69.81%	30.19%	**100.00%**
Total	**57.70%**	**42.30%**	**100.00%**

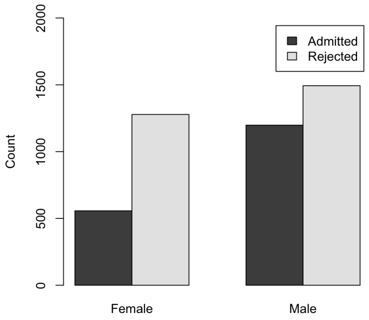
- **Question:** What can we take away from these observations?
- **Answer:**
  - Proportions that are _________ (____ or more) are clearly ___________
  - Proportions that are _______ require more _____________________ to make definitive conclusions

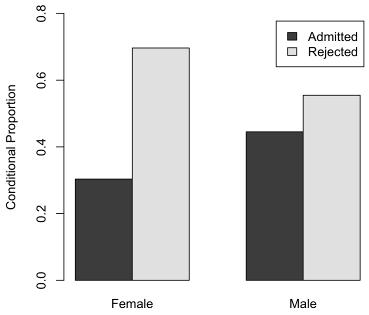
GRAPHICAL DISPLAYS OF TWO CATEGORICAL VARIABLES

- Each of the following displays the explanatory variable along the horizontal axis, but differs in how the response is displayed:
  - **Side-by-side bar graph:** a bar for each category of the response corresponding to the count or conditional proportion is created for each explanatory category
  - **Stacked bar graph:** a single bar for each explanatory category is divided into segments with the height of each segment corresponding to the conditional percentage in each group
  - **Mosaic plot:** a single bar for each explanatory category is divided into segments with the height corresponding to the conditional proportion in each group and the width corresponding to the number of cases in each explanatory category

Example: Side-by-Side Bar Graph

- **Scenario:** Comparing graduate admissions at Berkeley in 1973
- **Question:** What do the graphs tell you?

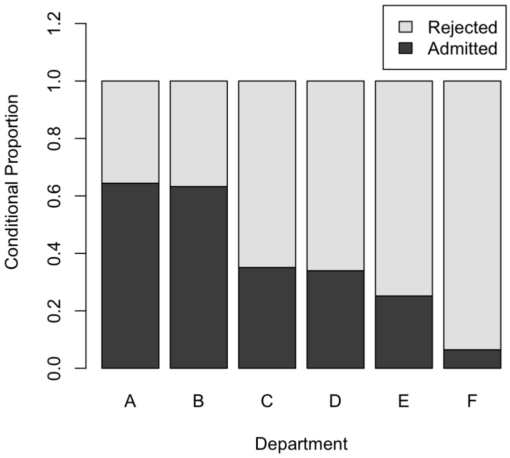




- **Answer:**
  - **Counts:** More males than females both __________ and ________________
  - **Proportions:** Sex and admission percentage are __________
    - About _____ of males were accepted compared to about __________ of females

Example: Segmented Bar Graph

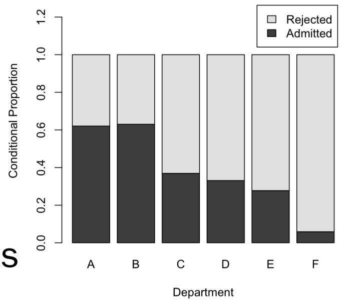
- **Scenario:** Comparing graduate admissions at Berkeley in 1973 across six departments regardless of sex
- **Question:** What can be said about the admission percentages in each department?
- **Answer:**
  - A and B had the __________ admission percentages (______)
  - C and D admitted about __________ of applicants while E was ______________
  - F had the ______ admission percentage (_______)



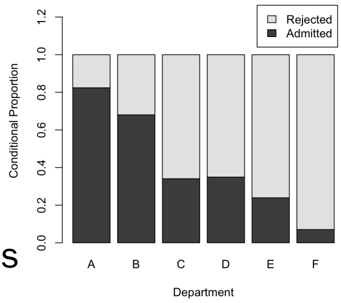
Example: Segmented Bar Graph

- **Scenario:** Comparing graduate admissions at Berkeley in 1973 across six departments, broken down by sex
- **Question:** What do you notice?
- **Answer:** Admission percentages within each department were ____________________
  - Females higher in ___________
  - Males slightly higher in ___________
- **Question:** What information is missing?
- **Answer:** _________________________ within each department

Males



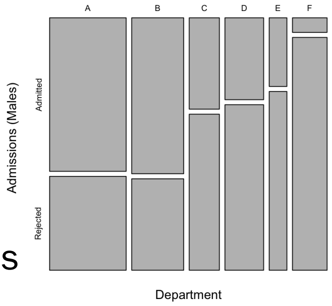
Females



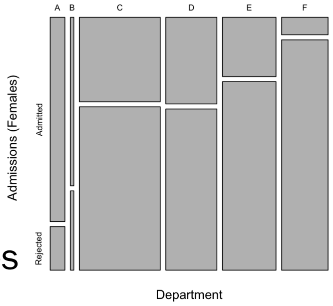
EXAMPLE: MOSAIC PLOT

- **Scenario:** Comparing graduate admissions at Berkeley in 1973 across six departments.
- **Question:** What do the mosaic plots reveal about where applicants applied?
- **Answer:** _______ of bar shows ____________ ___________ of applicants in each department
  - **A and B:** Received _____ of all male applications, but ______________ of female applications
  - **C, D, and E:** Got _____ of female applications, but only _____ of male applications
  - **F:** Conditional proportion of applicants was slightly _________________ (bar is _______)

Males



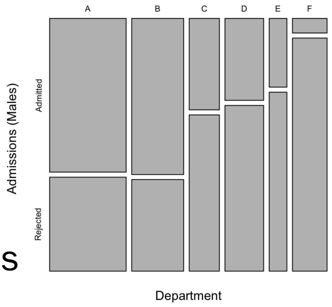
Females



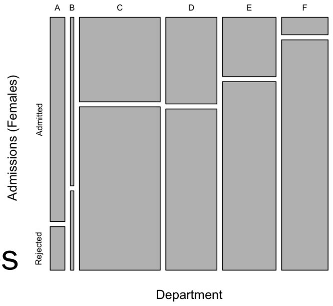
EXAMPLE: MOSAIC PLOT

- **Scenario:** Comparing graduate admissions at Berkeley in 1973 across six departments
- **Question:** How do the mosaic plots explain the apparent bias towards males?
- **Answer:** Males tended to apply to departments with ________________________ while females tended to apply to departments with ________________________
  - Resulted in overall admission percentage being driven ____ for males and driven ______ for females

Males



Females



SIMPSON'S PARADOX

- **Simpson's Paradox:** a phenomenon that occurs when there appears to be one trend in a set of data, but the trend disappears or reverses when the data is separated
- **Summarizing Berkeley Example:**
  - 44% of males were admitted compared to 35% of females overall, leading to claim of ______
  - Individual departments had ________________________, but admitted males and females at ________________
  - Number of applicants to each department skewed ______ admission rate ___ and _________ admission rate ________
  - Bias _____________ when separating data by _______________