Example:

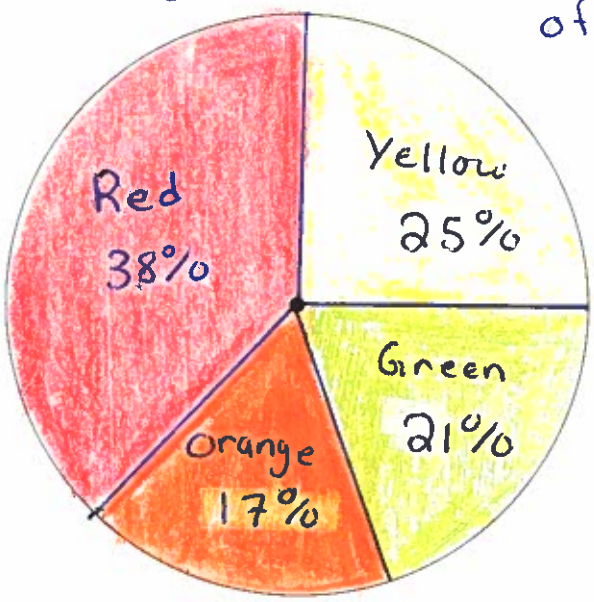
Construct a pie chart and bar charts for the candy example above.

Colour	Angle
Green	75°
Orange	60°
Red	135°
Yellow	90°

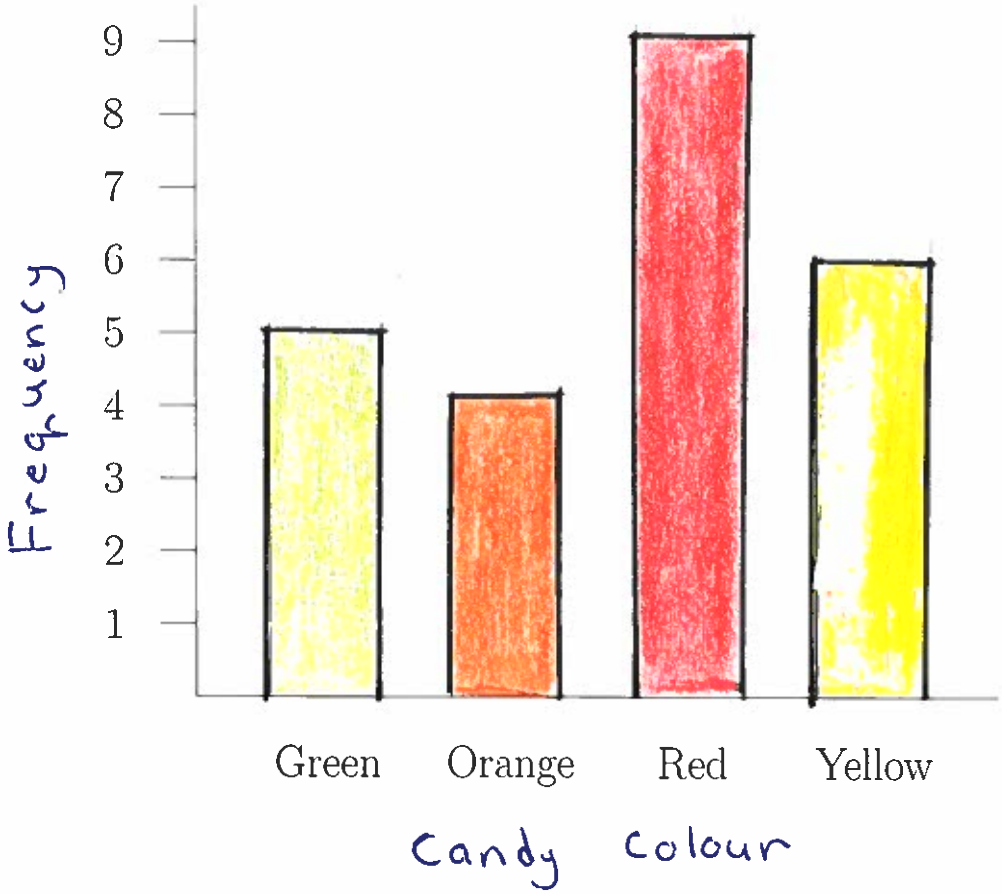
$\frac{5}{24}$ $\frac{4}{24}$ $\frac{9}{24}$ $\frac{6}{24}$

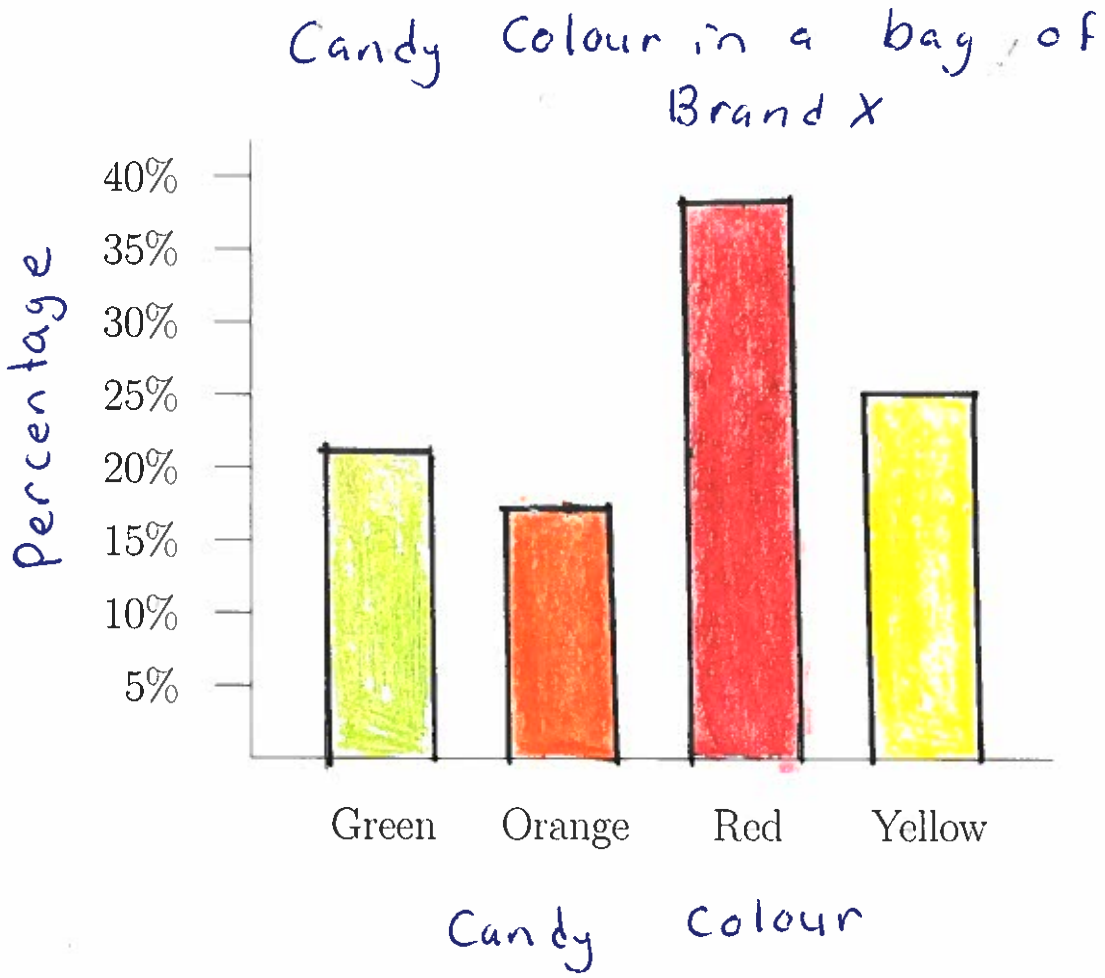
Green: $\frac{5}{24} \times 360° = 75°$

Candy Colours in a bag of Brand X



Candy Colours in a bag of Brand X




Candy Colour in a bag of Brand X
Percentage
40%
35%
30%
25%
20%
15%
10%
5%
Green
Orange
Red
Yellow
Candy Colour

Section 2.2: Exploring the Relationship Between Two Categorical Variables

Question: in studies in which we have collected data for two categorical variables, is there a relationship between the two variables? If so, how can we describe it?

association ≠ causation
↳ requires randomized experiments

Identify which of the two variables is the **explanatory** variable and which is the **response** variable.

comes first

The outcome of the response variable **depends on** or is **explained by** the explanatory variable.

Example: In a study to determine if smoking cigarettes affects blood sugar levels, smoking status is the explanatory variable and blood sugar level is the response variable.

Example:

- University student: number of trips home vs distance to home.

res exp

- T-shirts in a shop: price vs number sold.

exp res

- University graduate: choice of degree/major vs salary.

exp res

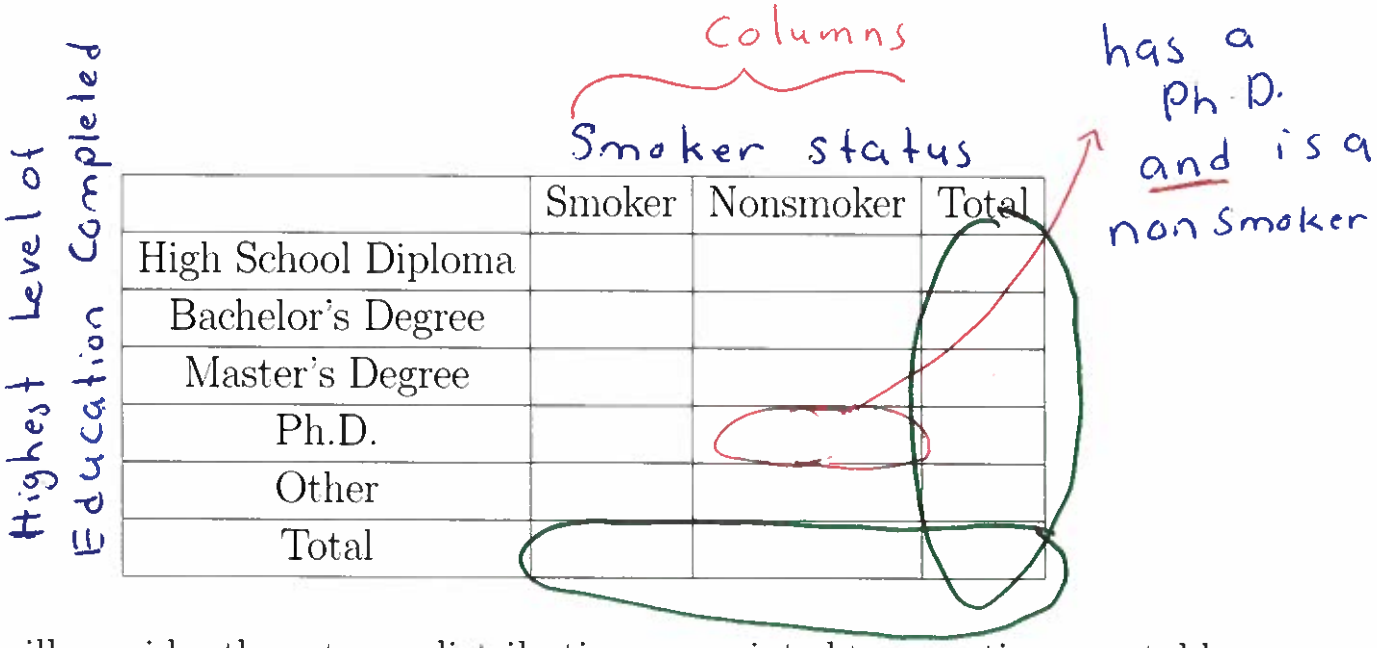
Warning: in observational studies Correlation does not imply causation! If there seems to be a relationship between two variables, it does **not** mean that one causes the other to occur. There may be a **lurking variable** (a variable which is hidden, but may be influencing our understanding of the relationship between the two variables).

"risk factors"

- use experiments to rule out lurking variables

Contingency Table: a two-way table which displays the frequencies of two categorical variables.

- the rows list the categories for one variable.
- the columns list the categories for the other variable.
- each entry on the table gives the number of observations that fall into both the row category and the column category for that entry.



	Smoker	Nonsmoker	Total
High School Diploma			
Bachelor's Degree			
Master's Degree			
Ph.D.			
Other			
Total			

We will consider three types distributions associated to a contingency table:

- **marginal distributions**.
- the **joint distribution**.
- **conditional distributions**.

The **margins** of a contingency table are the last row and last column of the table, which give the totals in each category of each variable. Each margin gives the frequency distribution of one of the variables. These distributions are called the **marginal distributions** of the contingency table.

The **joint distribution** is the distribution of both variables together in a contingency table, expressed as a percentage of the total number of observations. Each entry on the table gives the percentage of all observations that fall into both the row category and the column category for that entry.

Example: Sex vs. Liking Soccer

Liking Soccer Status

Sex	Likes	Dislikes	Total
Female	50	66	116
Male	44	32	76
Total	94	98	192

↓ % of total # of observations

joint distribution

	Likes	Dislikes	Total
Female	26%	34%	60%
Male	23%	17%	40%
Total	49%	51%	100%

Male and Likes: $\frac{44}{192} \times 100 = 23\%$

Marginal distribution for Liking Soccer Status

	Likes	Dislikes	Total
Frequency	94	98	192
Relative Frequency	0.49	0.51	1
Percentage	49%	51%	100%

% of row

Marginal distribution for Sex

	Frequency	Relative Frequency	Percentage
Female	116	0.6	60%
Male	76	0.4	40%
Total	192	1	100%

← % of column

The proportions of outcomes in each category of one variable that occur in each category of the other variable are called **conditional distributions**.

Original Question: To determine whether or not there is an association between the two variables, we examine the **conditional distribution of the response variable** for each of the categories of the explanatory variable. If the conditional distributions differ by a lot, then there is likely an association between the two variables. If there is no difference in the conditional distributions, then the two variables are said to be **independent**.

decide this? the same (no association)

Example: Sex vs. Liking Soccer:

Explanatory variable: Sex Response variable: liking soccer status

What are the conditional distributions of liking soccer by sex?

The conditional distribution of liking soccer for **females** is: 116 females

	Frequency	Relative Frequency	Percentage
Likes	50	0.43	43%
Dislikes	66	0.57	57%
Total	116	1	100%

The conditional distribution of liking soccer for **males** is: 76 males

	Frequency	Relative Frequency	Percentage
Likes	44	0.58	58%
Dislikes	32	0.42	42%
Total	76	1	100%

Conclusion: It appears that there is a difference in liking soccer between males and females. There may be an association between Sex and liking soccer. (Variables not independent)

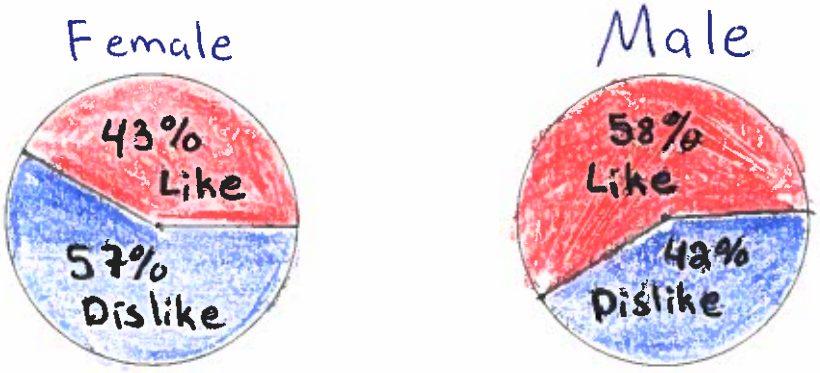
To visualize whether or not two variables are independent, we could use **side-by-side pie charts** (one pie chart for each category of the explanatory variable) or **segmented bar charts**. (A **segmented bar chart** shows the same information as a pie chart, but using bars instead of circles.)

↳ each bar totals 100%

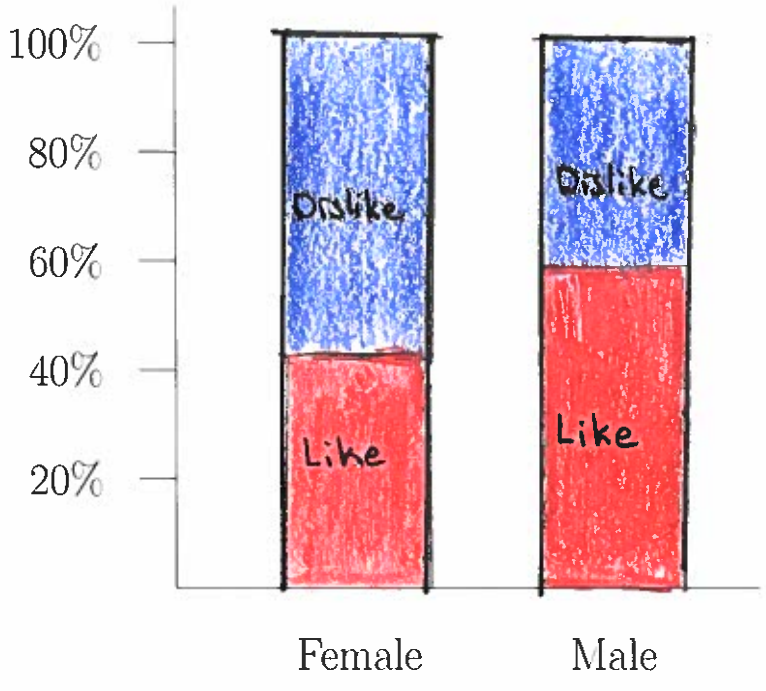
Example:

Female:	Likes soccer:	43%	Dislikes soccer:	57%
Male:	Likes soccer:	58%	Dislikes soccer:	42%

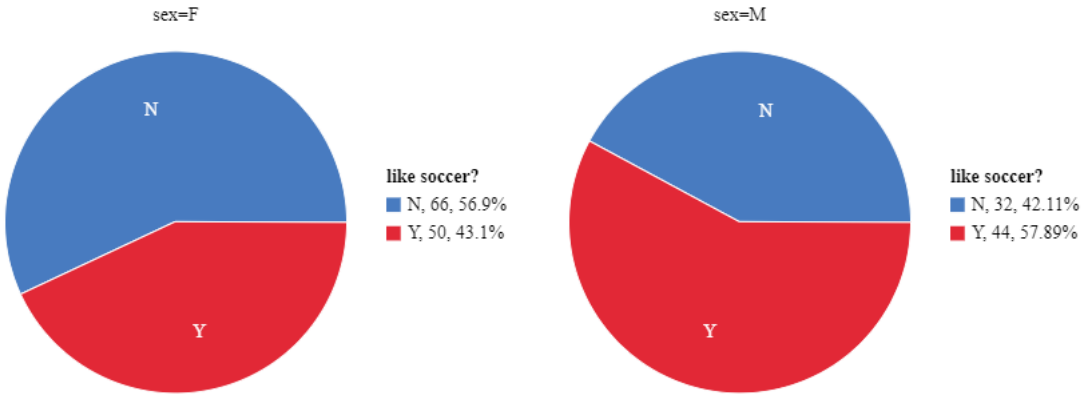
Side-by-side pie charts:



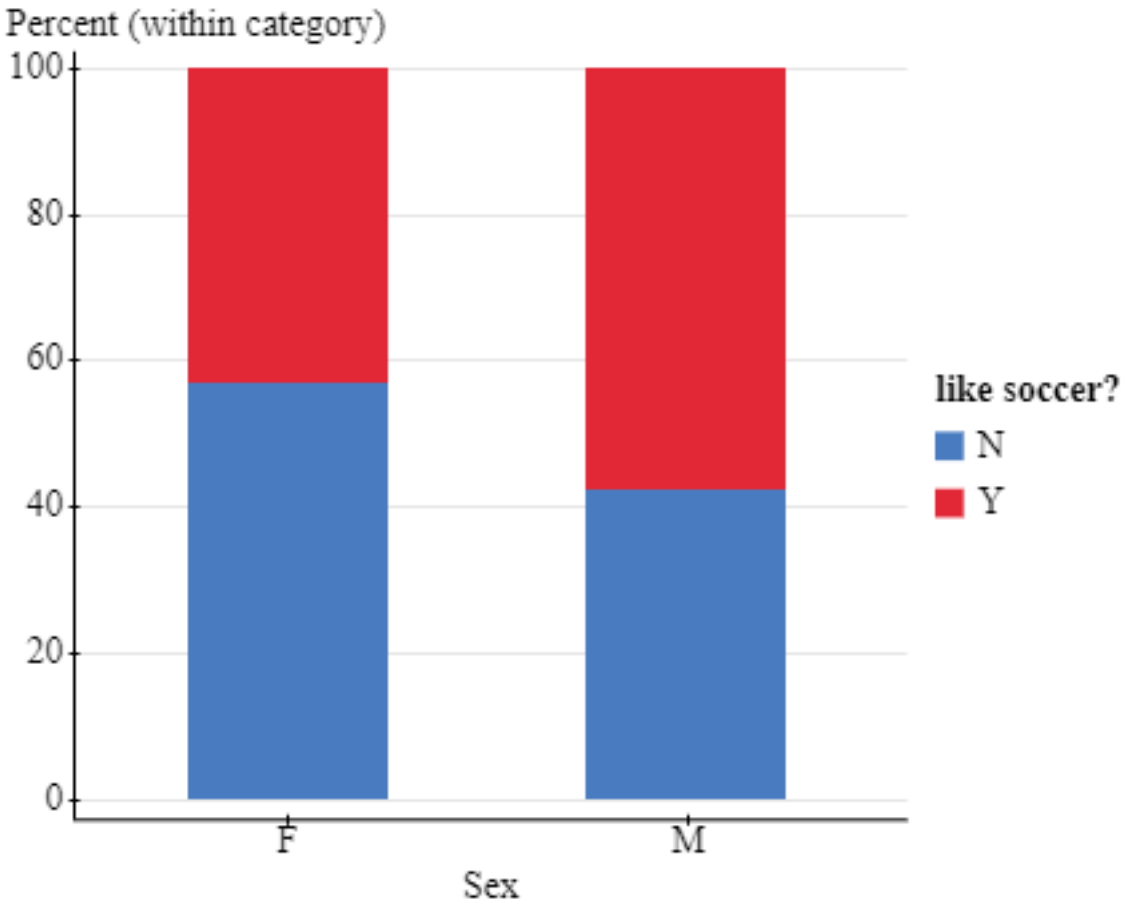
Segmented bar chart:



Side-by-side Pie Charts:



Segmented Bar Chart:



Note: The colours for each category in these charts are the same as the in-class version, but the order of the categories is different.

Example: Be careful with percentages:

	Likes	Dislikes	Total
Female	50	66	116
Male	44	32	76
Total	94	98	192

denominator

- What percent of the people are female?

$$\frac{116}{192} \times 100 \approx 60\%$$

- What percent of the people like soccer?

$$\frac{94}{192} \times 100 \simeq 49\%$$

- What percent of the females like soccer?

$$\frac{50}{116} \times 100 \simeq 43\%$$

- What percent of those who like soccer are female?

$$\frac{50}{94} \times 100 \simeq 53\%$$

- What percent of the people are female and like soccer?

$$\frac{50}{192} \times 100 \simeq 26\%$$

Chapter 3: Displaying and Summarizing Quantitative Data

To describe the distribution of a data set of a quantitative variable, we consider its:

- shape
- center
- spread

discuss together

Displaying Quantitative Variables with Graphs

Categorical	Quantitative
• Pie Charts	• Dotplots
• Bar Charts	• Stem-and-Leaf Displays
	• Histograms
	• Boxplots
	• Timeplots
	• Scatterplots

Dotplot: shows a dot for each observation, which is placed just above the value of that observation on the number line. The dots are stacked in a column over a value, so that the number of dots in the column represents the frequency of that value.

- effective for small data sets
- Shows data values

Winter 2020 Stat 151 Student Height