MATH 1342

Section 5.6

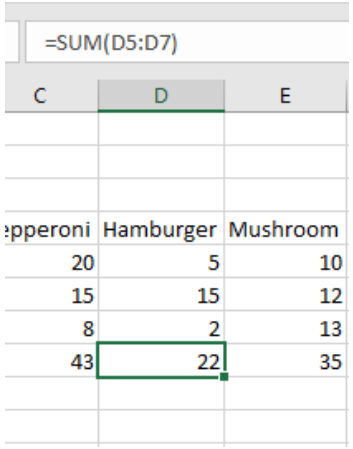
Relations in Categorical Data

A **two-way table** organizes the data for two categorical variables.

The totals of each row and column are considered **marginal distributions** because they appear in the margins of the table.

Marginal Distributions: Row and Column totals. These appear in the margins of the table.

Excel. Google sheets: http://sheets.google.com



Example:

The following two-way table describes the preferences in movies and pizza toppings for a random sample of 100 people.

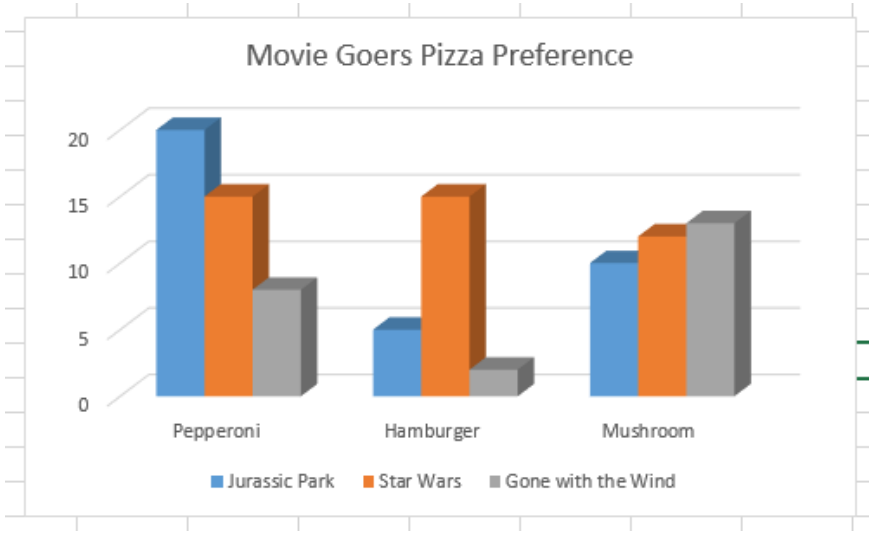
Movie	Pepperoni	Hamburger	Mushroom	
Jurassic Park	20	5	10	35
Star Wars	15	15	12	42
Gone with the Wind	8	2	13	23
	43	22	35	100

Enter the marginal distributions in the table.

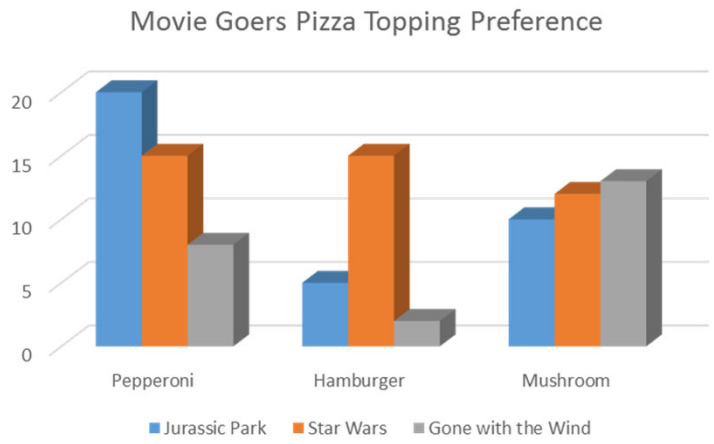
Draw a Bar Chart to Display the Marginal Distribution of Pizza Topping Preference

Using Excel would be the best option to do this.

Rstudio will work, but the syntax is very difficult to use.



Draw a Bar Chart to Display the Marginal Distribution of Pizza Topping Preference



What percent of our sample likes Gone with the Wind?

P(GwtW)= $\frac{23}{100} = .23$ or 23%

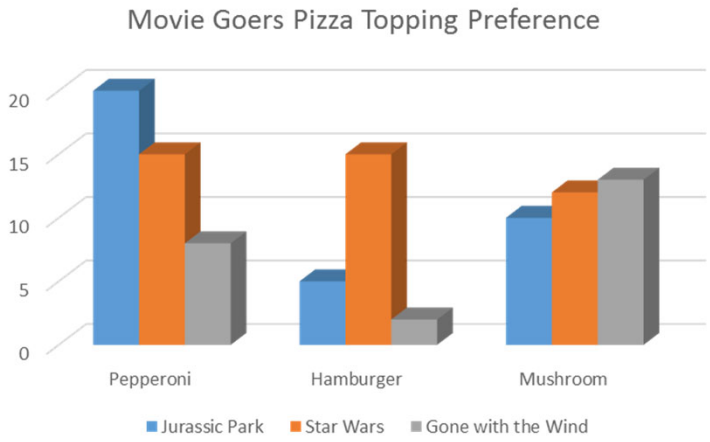
What percent of pepperoni lovers like Star Wars?

P(SW|Pepperoni)= $\frac{15}{43} = .3488$ or 35%

Movie	Pepperoni	Hamburger	Mushroom	
Jurassic Park	20	5	10	35
Star Wars	15	15	12	42
Gone with the Wind	8	2	13	23
	43	22	35	100

Draw a Bar Chart to Display the Marginal Distribution of Pizza Topping Preference

Movie	Pepperoni	Hamburger	Mushrooms	
Jurassic Park	20	5	10	
Star Wars	15	15	12	
Gone with the Wind	8	2	13	



A **conditional distribution** is made up of the percentages that satisfy a given condition.

$$P(A|B) = \frac{n(A \cap B)}{n(B)}$$

Movie	Pepperoni	Hamburger	Mushroom	
Jurassic Park	20	5	10	35
Star Wars	15	15	12	42
Gone with the Wind	8	2	13	23
	43	22	35	100

Popper 17:

Compare the conditional distributions of movie preference for hamburger lovers and mushroom lovers. Back up your description with percentages.

1. What percent of hamburger lovers like Star Wars? $\frac{15}{22} = .6818$

a. 23% b. 36% c. 68% d. 15%

2. What percent of hamburger lovers like Jurassic Park? $\frac{5}{22} = .2272$

a. 23% b. 12% c. 5% d. 35%

3. What percent of mushroom lovers like Gone with the Wind?

a. 59% b. 37% c. 2% d. 22% $\frac{13}{35} = .3714$

Simpson's Paradox

Always be careful if combining data to make a comparison. **Simpson's Paradox** is the reversal of the direction of a comparison or an association when data from several groups are combined to form a single group.

This a way of being deceitful when presenting data without falsifying anything.

Be mindful of this when looking at other people's statistics.

Simpson's Paradox is regrouping of raw in such a way that the results appear different.

In a 1991 study by Radelet and Pierce of the effect of race on death-penalty sentences, the following table was obtained tabulating the death-penalty sentences (Death) and non-death-penalty sentences (No death) in murder convictions in the state of Florida.

Defendant's race	Death	No death	Percent death
Caucasian	53	430	11.0
African-American	15	176	7.9

Now, we consider *the very same data*, except that we stratify according to the **race of the victim** of the murder. Below is the table.

Victim's race	Defendant's race	Death	No death	Percent death
Caucasian	Caucasian	53	414	11.3
Caucasian	African-American	11	37	22.9
African-American	Caucasian	0	16	0.0
African-American	African-American	4	139	2.8

Here we see that when considering the cases involving Caucasian victims separately from the cases involving African-American victims, that the African-American *defendants* are more likely than Caucasian ones to receive the death penalty in both instances (22.9% vs 11.3% in the first case and 2.8% vs. 0.0% in the second case).

- This is adapted from Subsection 2.3.2 of A. Agresti (2002), *Categorical Data Analysis*, 2nd ed., Wiley, pp. 48-51.