**Methods for Association Between Two Categorical Variables**

**Example 3.1: MythBusters and the Yawning Experiment**

MythBusters, a popular television program on the Discovery Channel, once conducted an experiment to investigate whether or not yawning is contagious. The premise of the experiment was to invite a stranger to sit in a booth for an extended period of time. Fifty subjects were said to be tested in total, of which 34 were "seeded" with a yawn by the person conducting the experiment. The other 16 were not given a yawn seed. Using a two-way mirror and a hidden camera, the experimenters observed and recorded the results which are given below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yawned | Did not Yawn | Total |
| Seeded with a yawn | 10 | 24 | 34 |
| Not seeded with a yawn | 4 | 12 | 16 |

**Research Hypothesis: Those “seeded” with a yawn are more likely to actually yawn than those who are not seeded.**

When we analyze data on two variables, our first step is to distinguish between the **response** variable and the **explanatory** (or **predictor**) variable.

|  |
| --- |
| **Response variable:** The outcome variable on which comparisons are made.  **Explanatory (or Predictor) variable:**  This defines the groups to be compared. |

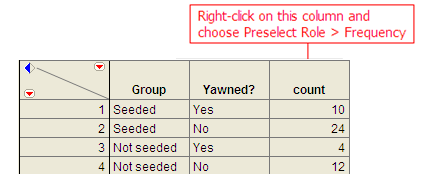
Questions:

1. What variables must be measured in the MythBusters Yawning experiment in order to investigate the research hypothesis? Are these variables categorical or numerical?

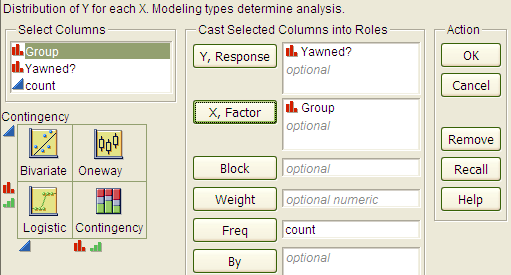
1. Which is the response variable? Which is the explanatory variable?

|  |
| --- |
| **Descriptive Methods for Two Categorical Variables:**  A **contingency table** shows the joint frequencies of two categorical variables. The rows of the table denote the categories of the first variable, and the columns denote the categories of the second variable.  A **mosaic plot** gives a visual representation of the relationship between two categorical variables. A mosaic plot graphically presents the information given in the contingency table. |

To obtain these descriptive summaries from JMP, the data should be entered as follows:



Select **Analyze > Fit Y by X**. Place the response variable in the **Y, Response** box and the explanatory variable in the **X, Factor** box.



Click OK, and JMP returns the following:

|  |  |
| --- | --- |
|  |  |

Questions:

1. Find the proportion that yawned in the Seeded group.
2. Find the proportion that yawned in the Not Seeded group.
3. Find the difference in the proportion that yawned between these two groups. Do these proportions differ in the direction conjectured by the researchers?
4. Even if the seeding of a yawn had absolutely no effect on whether or not a subject actually yawned, is it *possible* to have obtained a difference such as this by random chance alone? Explain.

**Visualizing No Association**

When interpreting a contingency table and mosaic plot, it is important to be able to visualize what these table would look like if there was no connection between the variables. To get a feel for how to construct the *no association table* and *plot,* let’s consider two variables that are clearly unrelated.

1. **Variable 1:** Whether or not I am wearing a blue shirt.
2. **Variable 2:** Whether or not it is cloudy.

Suppose, for the sake of this example, I wear blue shirts 68% of the time. Consider the following contingency table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cloudy | Not Cloudy | Total | Proportion |
| Blue Shirt |  | 36\*0.68=24.48 | 34 | 34/50 = 0.68 |
| Shirt of Another Color |  | 36\*0.32 = 11.52 | 16 | 16/50 = 0.32 |
| Total | 14 | 36 | 50 |  |

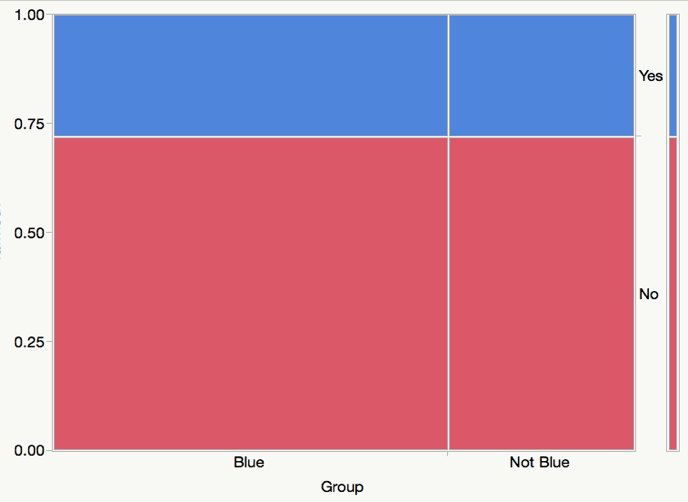
Questions:

1. Note that we have illustrated how to compute the proportion of days with a shirt of another color. Compute the proportion of the 50 observations had a blue shirt.
2. Note that 32% of days are “not blue days.” If there is no association between my shirt color and the weather, argue that 32% of the days that are labeled “not cloudy” should also be labeled “not blue”.

In total this would be 36\*0.32 = 11.52 days on average. This value is known as the **expected count assuming no association** or simply the **expected count**. There is a nice short cut for computing the values in the no association table. Notice that the 16\*36/50 = 11.52. That is we can compute the expected value given no association by multiplying the row and column totals, then dividing by the grand total.

1. Fill out the rest of the table, assuming there is no connection between my shirt color and whether or not it is cloudy.

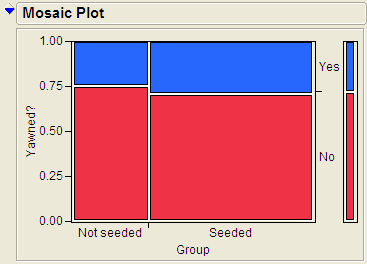
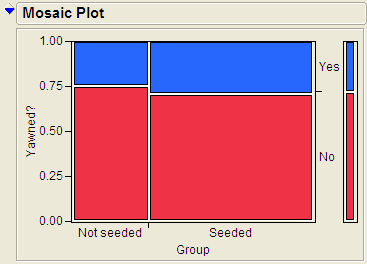
Let’s look at the mosaic plot for the no association table.



Notice how the bars line up perfectly with the short bars on the right (which represents to marginal distribution of Yes and No). This will always be approximately true if there is little or no association between the two variables. In this way, we can use the marginal distribution on the right to quickly visualize the *no association* plot.

Let’s return to the original example and illustrate this visualization.

1. Start with the original mosaic plot
2. Draw a line straight across from the marginal distribution.

1. Comment on how the actual plot differs from the no association plot. Which groups have more or less value than expected?
2. Based on this graph, do think there is an association between whether or not a person yawned and whether or not they were seeded with a yawn? Explain.

**Example 3.2: Vested Interest and Task Performance** *This example is from Investigating Statistical Concepts, Applications, and Methods by Beth Chance and Allan Rossman. 2006. Thomson-Brooks/Cole.*

“A study published in the *Journal of Personality and Social Psychology* (Butler and Baumeister, 1998) investigated a conjecture that having an observer with a vested interest would decrease subjects’ performance on a skill-based task. Subjects were given time to practice playing a video game that required them to navigate an obstacle course as quickly as possible. They were then told to play the game one final time with an observer present. Subjects were randomly assigned to one of two groups. One group (A) was told that the participant and observer would each win $3 if the participant beat a certain threshold time, and the other group (B) was told only that the participant would win the prize if the threshold were beaten. The threshold was chosen to be a time that they beat in 30% of their practice turns. The following results are very similar to those found in the experiment: 3 of the 12 subjects in group A beat the threshold, and 8 of 12 subjects in group B achieved success.”

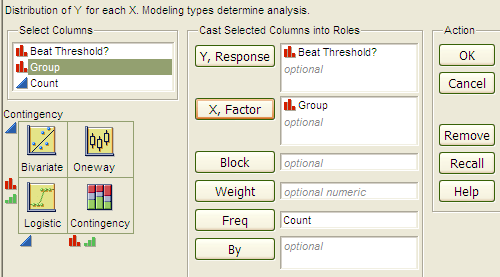
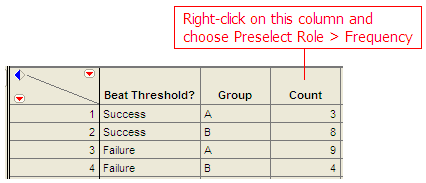
|  |  |  |  |
| --- | --- | --- | --- |
|  | A: Vested Interest | B: No Vested Interest | Total |
| Achieved success | 3 | 8 | 11 |
| Did not achieve success | 9 | 4 | 13 |
| Total | 12 | 12 | 24 |

**Research Hypothesis: Having an observer with a vested interest decreases performance on a skill-based task.**

Questions:

1. What variables must be measured in order to address the research question? Are they categorical or numerical?
2. Which is the response variable? Which is the explanatory variable?

The data can be entered into JMP as follows:



To create the contingency table and the mosaic plot in JMP, select **Analyze > Fit Y by X**. As always, place the response variable in the **Y, Response** box and the explanatory variable in the **X, Factor** box. Click OK, and JMP returns the following:

|  |  |
| --- | --- |
|  |  |

Questions:

1. What is the proportion of successes (beating the threshold) for each group?
2. Add a line to the mosaic plot that illustrates how this graph would look if there was no association between the group and beat threshold.
3. Which groups are over-represented when compared to the no association graph. How different are these graphs?
4. Even if the observer’s interest had absolutely no effect on subjects’ performance, is it *likely* to have obtained a difference such as this by random chance alone? Do you think this difference provides strong evidence that there is an association between achievement and whether or not the participant has a vest interest? Explain.
5. Even if the observer’s interest had absolutely no effect on subjects’ performance, is it *possible* to have obtained a difference such as this by random chance alone? Explain.

**Example 3.11: Support for Iraq War and Political Affiliation**

In March of 2003, the Pew Internet & American Life Project commissioned Princeton Survey Research Associates to develop and carry out a survey of what Americans thought about the recent war in Iraq. Some of the results of this survey of over 1,400 American adults are given in the JMP data file **IraqWar.JMP**. (*Source:* *McClave & Sincich, Problem 13.33*)  
  
Responses to the following questions were recorded:

1. Do you support or oppose the Iraq War?
2. Do you ever go online to access the Internet or World Wide Web?
3. Do you consider yourself a Republican, Democrat, or Independent?
4. In general, would you describe your political views as very conservative, conservative, moderate, liberal, or very liberal?
5. What is your race?
6. Do you live in a suburban, rural, or urban community?

**Research Question: Is there a significant association between Support for the Iraq War and Political Affiliation?**

Note that this investigation requires us to focus on only two of the measured variables: *Support for the Iraq War* and *Political Affiliation*. First, let’s summarize the data using JMP, which returns the contingency table and mosaic plot:

|  |  |
| --- | --- |
|  |  |

Questions:

1. Draw a line to illustrate the no association graph, that is the mosaic plot if there was no association between political affiliation and support for the war.
2. Identify and comment on the combinations of labels that are over-represented, when compared to the no association plot.
3. What can you say about the association between Support for the Iraq War and Political Affiliation based on the data obtained in the sample? Do you think there is evidence of an association between political affiliation and support for the war? Explain.