In this handout, we will continue to discuss ways to investigate the relationship between two categorical variables. Recall that a categorical variable takes on a value that is the name of a category or a label. When examining the relationships between variables, it may also help to consider the following definitions.

**Definitions**

**Dependent Variable:** The variable whose changes we wish to study. This is often called the *response* variable.

**Independent Variable:** A variable under consideration that is used to explain changes in the dependent variable. This is often called the *predictor* variable. Note that studies often involve more than one independent variable.

**Example: Investigating Survival on the Titanic**

Consider the following screenshot of data from the Titanic, the famous ship that sunk in 1912. This dataset can be found in the *Titanic.xlsx* file on the course web site.

Suppose you were asked to investigate the following questions. Identify the independent and dependent variable for each scenario.

1. Are 1st class passengers more likely to have survived than 3rd class passengers?
   - **Dependent Variable** (i.e., the response): *Survival Status*
   - **Independent Variable** (i.e., the predictor): *Passenger Class*
2. Are females more likely to have survived than males?

- Dependent Variable (i.e., the response): **Survival Status**
- Independent Variable (i.e., the predictor): **Gender**

An image of the R.M.S. Titanic floorplan.

Our task is to investigate the relationship between Survival and other variables such as Passenger Class and Gender. Ultimately, we want to get a table of counts like we did in Handout 7. Then, we will calculate relevant percentages from this table to help us investigate the questions of interest. The Excel feature that will allow us to obtain appropriate summaries is called a **PivotTable**.
PivotTables are one of the most under-utilized tools in Excel when it comes to analyzing data. PivotTables are powerful and permit us to “slice” the data in various ways, which is important when investigating relationships between variables.

**Getting a PivotTable in Excel**

To obtain a PivotTable in Excel, first highlight the data that you want to summarize. Then, select **Insert > PivotTable** as is shown here.

The Create PivotTable dialog box will appear. If you have previously highlighted the data that you want to summarize, it will appear in the “Select a table or range” box. If not, you should enter the range of the data to be summarized in this box before proceeding.

The radio button near the bottom allows you to specify where the PivotTable will be placed. I typically place the PivotTable on a New Worksheet. Once you have selected this, click OK.
The following is returned by Excel. This template is the basic template for a PivotTable.

To create a PivotTable, simply select the variables you’d like to summarize and place them in either a row or column. The independent variable (i.e., the predictor) is typically placed in the rows of the table and the dependent variable (i.e., the response) in the columns of the table.
Notice that this created Row and Column Labels for a table in the Excel sheet; however, nothing has been summarized in the table so far. To obtain the appropriate counts in the table, you must drop one of the variables of interest into the Values box in the lower right-hand corner.

To summarize, this is how we would set up the PivotTable in Excel to investigate differences in Survival Status across Passenger Class.

- The **dependent** variable (i.e., Survived) was placed in the Column Labels box
- The **independent** variable (i.e., Passenger Class) was placed in the Row Labels box
- Survived was placed in the Values box to obtain the **counts**

Do this in Excel and use your PivotTable to obtain the appropriate counts. Write them into the following table. In the previous handout, we called this a bivariate frequency table. Statisticians, however, often refer to this type of summary as a **contingency table** or cross-tab table (cross-tab is short for cross-tabular).
### Questions

1. How many first class passengers were on the Titanic? How many 3rd class passengers?

   1st class: 323  
   3rd class: 709

2. How many passengers survived? Did a majority of the passengers survive?

   500 of 1309 passengers survived (38.2%). The majority did not survive.

3. Your friend makes the following statement: ”More 3rd class passengers survived than 2nd class passengers (181 vs. 119).” This statement is technically true, but it is misleading. Explain why this is misleading.

   It is misleading because it fails to take into consideration the total number of 2nd and 3rd class passengers on the trip.

Recall that when making comparisons between groups, we must take into consideration that the groups may be of different sizes. To alleviate this concern, we compute row percentages.

### Questions

4. Consider only the 3rd class passengers. How many 3rd class passengers were there? How many survived?

   181 of 709 3rd class passengers survived
5. What percentage of the 3rd class passengers survived?

\[ \frac{181}{709} = 25.5\% \]

6. Consider next the 2nd class passengers. What percentage of the 2nd class passengers survived?

\[ \frac{119}{277} = 43\% \]

7. Compare the percentage of 3rd class passengers that survived to the percentage of 2nd class passengers that survived. How different are these percentages?

The survival rate was higher for the 2nd class passengers.

8. Suppose your friend disagrees with how you computed the survival rate for each group above and instead uses the total number of passengers in the denominator.

Survival Rate for 2nd class passengers: \( \frac{119}{1309} \approx 9\% \)
Survival Rate for 3rd class passengers: \( \frac{181}{1309} \approx 14\% \)

Compare these percentages.

These incorrect calculations make it appear as though the survival rate was higher for the 3rd class passengers.

9. Do the comparisons made in Question 7 agree with those in Question 8? If they differ, which comparisons are more meaningful? Discuss.

They disagree, but the correct comparisons were made in Question 7.
Getting Row Percentages in Excel

Excel can be used to directly calculate the necessary row percentages so that fair comparisons can be made across Passenger Class. To obtain these row percentages, right-click on the Count of Survived cell in your PivotTable. Select Value Field Settings…

In the Value Field Settings window, select the Show Values As tab, and select % of Row Total from the Show values as drop-down menu. Click OK.

Use your PivotTable output to obtain the missing row percentages in the following table.

<table>
<thead>
<tr>
<th>Passenger Class</th>
<th>Survived</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>1st Class</td>
<td>38.1%</td>
<td>61.9%</td>
</tr>
<tr>
<td>2nd Class</td>
<td>57.0%</td>
<td>43%</td>
</tr>
<tr>
<td>3rd Class</td>
<td>74.5%</td>
<td>25.5%</td>
</tr>
</tbody>
</table>
Questions

10. Explain how the value 61.9% was computed in the above table. Show the formula with the actual numbers.

\[ \frac{200}{323} = 61.9\% \]

11. In this example, the percentages in the first row add up to 100%. Will this always be the case when row percentages are computed? Discuss.

Yes.

12. Compare the survival rate of 1\textsuperscript{st} class passengers to that of 3\textsuperscript{rd} class passengers. How different are these survival rates?

The survival rate was much higher for 1\textsuperscript{st} class passengers than for 3\textsuperscript{rd} class passengers.

The following graph shows the survival rate (i.e., the percentage that survived) for each passenger class.

Questions

13. What is the exact height of the 1\textsuperscript{st} Class bar? That is, what values were used to construct this plot?

61.9\%
14. What pattern(s) do you see in the plot? Discuss. **The survival rate decreases from 1st to 2nd to 3rd class.**

When Excel is used to create a chart based on the PivotTable results, it includes all of the row percentages in the chart – not just the row percentages for those that survived.

![Passenger Class vs. Survival Chart](chart.png)

**Questions:**

15. On the above plot, sketch the appropriate bar for the proportion that died for the 2nd class and 3rd class passengers. What pattern(s) is/are present when the proportion that died is considered across the passengers classes? Discuss.

**The death rate increases from 1st to 2nd to 3rd class.**

The graph shown above is typically called a side-by-side bar chart (Excel calls it a clustered column chart). Such charts are used very often to summarize relationships between two categorical variables, but we can do better than this. One suggestion to improve this graphical summary of the relationship between Passenger Class and Survival is to stack the bars for each passenger class instead of using the side-by-side bars.

![Passenger Class vs. Survival Chart](chart.png)
Creating 100% Stacked Column Charts in Excel

To create this graph in Excel, first highlight the cells containing your PivotTable. Then, select Insert > Column Chart > 100% Stacked Column.

Excel returns the following:
To remove some of the white space between the columns, select one of the bars, right-click, and select Format Data Series… In the Format Data Series window, you can reduce the Gap width. You can also right-click and “Add Data Labels” to display the appropriate row percentages on your chart. Finally, you can add a title if desired using the Chart Layouts menu.

The resulting chart is shown below:

![Chart showing Survival Status Across Passenger Class]

**Investigating Survival Status Across Gender**

Next, consider the relationship between Survival and Gender.

Tasks:

1. Use the PivotTable feature in Excel to obtain the following table.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Survived</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Female</td>
<td>127</td>
<td>339</td>
</tr>
<tr>
<td>Male</td>
<td>682</td>
<td>161</td>
</tr>
<tr>
<td>Total</td>
<td>809</td>
<td>500</td>
</tr>
</tbody>
</table>
2. Obtain the appropriate row percentages using Excel. Write these in the table below

<table>
<thead>
<tr>
<th>Gender</th>
<th>Survived</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>Total</td>
</tr>
<tr>
<td>Female</td>
<td>27.25%</td>
<td>72.75%</td>
<td>100%</td>
</tr>
<tr>
<td>Male</td>
<td>80.9%</td>
<td>19.1%</td>
<td>100%</td>
</tr>
<tr>
<td>Total</td>
<td>61.8%</td>
<td>31.2%</td>
<td>100%</td>
</tr>
</tbody>
</table>

3. Create a 100% Stacked Column Chart using the row percentages computed above. Sketch in your result below.

Questions:

1. What was the survival rate of Females on the Titanic?

   **72.75%**

2. What was the survival rate of Males on the Titanic?

   **19.10%**

3. How different were the survival rates across Gender? Discuss.

   **Females were much more likely to survive.**
Considering a Third Variable

The previous investigation considered the relationship between Survival and Gender. What happens if we also consider whether the passenger was a child or an adult? To investigate this, simply drag the Child/Adult variable to the Report Filter box.

Using the Filter in either the Pivot table or chart, you can select to see information for the children and adults separately. Do this to complete the following table:

<table>
<thead>
<tr>
<th>Survival Rates</th>
<th>Adult</th>
<th>Child</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>73.35%</td>
<td>17.08%</td>
</tr>
</tbody>
</table>

You can also examine the relationships between variables in the following plots.
Questions:

4. Does the relationship between Gender and Survival change based on whether the passenger was a child or an adult? Explain.

The difference in survival rates across gender was larger for adults than for children. In both cases, males were more likely to die than females; for adults, however, the difference in death rates between males and females was more pronounced.

Example: Political Affiliation and Opinions on Immigration Reform

Recall the example from Handout 7. A recent CNN/ORC opinion poll surveyed a random sample of 1,010 adults nationwide. Respondents were asked the following question: “What should be the main focus of the U.S. government in dealing with the issue of illegal immigration: developing a plan that would allow illegal immigrants who have jobs to become legal U.S. residents, or developing a plan for stopping the flow of illegal immigrants into the U.S. and for deporting those already here?”

The data from this opinion poll are reported in the following contingency table:

<table>
<thead>
<tr>
<th></th>
<th>Legal residency</th>
<th>Stop flow/deport</th>
<th>Unsure</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat</td>
<td>202</td>
<td>79</td>
<td>12</td>
<td>293</td>
</tr>
<tr>
<td>Independent</td>
<td>267</td>
<td>189</td>
<td>29</td>
<td>485</td>
</tr>
<tr>
<td>Republican</td>
<td>79</td>
<td>144</td>
<td>9</td>
<td>232</td>
</tr>
<tr>
<td>Totals</td>
<td>548</td>
<td>412</td>
<td>50</td>
<td>1,010</td>
</tr>
</tbody>
</table>

Suppose that the question of interest is as follows: Do opinions on this issue tend to differ across political affiliation?

Questions:

1. What are the two variables of interest in this study? Political Affiliation and Opinion on Immigration Reform

2. Which is the dependent variable (i.e., the response)? Opinion on Immigration Reform

3. Which is the independent variable (i.e., the predictor)? Political Affiliation
Analyzing Data from an Existing Contingency Table in Excel

In Handout 7, we computed percentages that would be relevant for investigating this question of interest. Here, our objective is to compute these percentages and to obtain a graphical summary of the data using Excel.

Note that in terms of how the data were presented to us, there is one big difference between this example and the Titanic example. In the Titanic example, our Excel file contained the raw data. In other words, each row in the spreadsheet represented a single subject in the study.

We were given the raw data for the Titanic example, and we had to create a contingency table using PivotTables in Excel. We were not provided with raw data for the Immigration Reform study; instead, the data were given to us in the form of a contingency table. If we were to use the exact same steps in Excel for the Immigration Reform example, we would need to create an Excel sheet that was set up so that each row represents the outcomes for a single subject in the study. For example, we could start with something like this:

This is just a start. There were 1,010 respondents in this study, so we would have to create a sheet with 1,010 rows. Of these rows, 202 would identify Democrats who said “Legal residency,” 79 would identify Democrats who said “Stop flow/deport,” etc. Doesn’t sound like much fun, does it? Instead of going through all of this trouble, let’s take a short-cut. In Excel, we can enter the data as follows. Essentially, we will give each cell in the contingency table its own row in Excel. Then, we’ll add a column to our data sheet that tells Excel how many respondents appeared in each cell.
Enter the data from the contingency table in Excel as shown below.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Political Affiliation</td>
<td>Opinion</td>
<td>Count</td>
</tr>
<tr>
<td>2</td>
<td>Democrat</td>
<td>Legal residency</td>
<td>202</td>
</tr>
<tr>
<td>3</td>
<td>Independent</td>
<td>Legal residency</td>
<td>257</td>
</tr>
<tr>
<td>4</td>
<td>Republican</td>
<td>Legal residency</td>
<td>79</td>
</tr>
<tr>
<td>5</td>
<td>Democrat</td>
<td>Stop flow/deport</td>
<td>79</td>
</tr>
<tr>
<td>6</td>
<td>Independent</td>
<td>Stop flow/deport</td>
<td>189</td>
</tr>
<tr>
<td>7</td>
<td>Republican</td>
<td>Stop flow/deport</td>
<td>144</td>
</tr>
<tr>
<td>8</td>
<td>Democrat</td>
<td>Unsure</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>Independent</td>
<td>Unsure</td>
<td>29</td>
</tr>
<tr>
<td>10</td>
<td>Republican</td>
<td>Unsure</td>
<td>9</td>
</tr>
</tbody>
</table>

Now, we can create a PivotTable from these data.

- Just like we did with the Titanic example, highlight the cells containing the data, and insert a PivotTable. Be sure to place the independent variable in the rows and the dependent variable in the columns.

- Here is the difference between this example and the Titanic example: when we are working with data that have already been summarized with counts in Excel (i.e., when each row in the Excel file represents a cell in a contingency table instead of a single subject in the study), then place the **Count** in the Values box.
You should see the following PivotTable.

<table>
<thead>
<tr>
<th>Row Labels</th>
<th>Column Labels</th>
<th>Legal residency</th>
<th>Stop flow/deport</th>
<th>Unsure</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat</td>
<td>202</td>
<td>79</td>
<td>12</td>
<td>293</td>
<td></td>
</tr>
<tr>
<td>Independent</td>
<td>267</td>
<td>189</td>
<td>29</td>
<td>485</td>
<td></td>
</tr>
<tr>
<td>Republican</td>
<td>79</td>
<td>144</td>
<td>9</td>
<td>232</td>
<td></td>
</tr>
<tr>
<td>Grand Total</td>
<td>548</td>
<td>412</td>
<td>50</td>
<td>1010</td>
<td></td>
</tr>
</tbody>
</table>

Just like we did with the Titanic Data, we can look at relevant Row Percentages instead of counts:

<table>
<thead>
<tr>
<th>Row Labels</th>
<th>Column Labels</th>
<th>Legal residency</th>
<th>Stop flow/deport</th>
<th>Unsure</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat</td>
<td>68.94%</td>
<td>26.96%</td>
<td>4.10%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td>Independent</td>
<td>55.05%</td>
<td>38.97%</td>
<td>5.98%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td>Republican</td>
<td>34.05%</td>
<td>62.07%</td>
<td>3.88%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td>Grand Total</td>
<td>54.26%</td>
<td>40.79%</td>
<td>4.95%</td>
<td>100.00%</td>
<td></td>
</tr>
</tbody>
</table>

Finally, we can create the 100% stacked column chart. Do this in Excel, and sketch the results below.

Questions:

4. Use the information in this plot to write a sentence or two addressing the question of interest in this study: Do opinions on this issue tend to differ across political affiliation? Democrats in this sample were more likely to answer “legal residency” while Republicans were more likely to answer “stop flow/deport.”