Chapter 4: Understanding Relationships between Two Categorical Variables

Material presented in this chapter will allow us to investigate, measure, and test relationships between *two* categorical variables. Relationships between variables or outcomes may be complicated and may involve more than two variables (e.g. the risk of heart disease differs between the genders and across race); however, these types of investigations are beyond the scope of this class.

|  |
| --- |
| Definition |
| Response Variable: The primary variable or outcome of interest in the investigation  Predictor Variable: The other variable(s) being considered in the investigation |

If the relationship between two variables is being studied, then it may be beneficial to label these variables. The labels given to the two variables are response variable and predictor variable.

Consider once again the data presented in Chapter 3 regarding the Minnesota Student Survey.



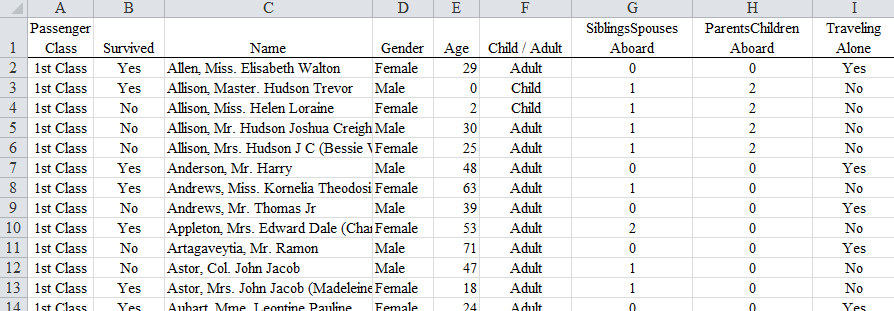
Identify the two variable types for each of the following graphs.

|  |  |
| --- | --- |
| Relationship Between Survey Outcome and Gender for Grade 12 | Relationship Between  Survey Outcome and Grade |
| Response Variable:  Predictor Variable: | Response Variable:  Predictor Variable: |

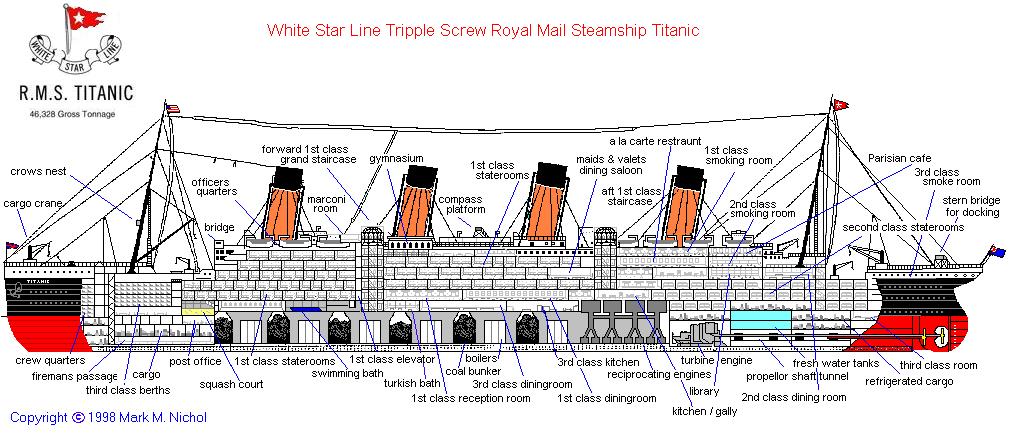
4.1: Investigating the Relationship

In this section, we will consider how to summarize, measure, and visualize the relationship between two categorical variables.

Example 4.1.1: Consider the following snip-it of data from the Titanic, the famous ship that sunk in 1912. This dataset can be found in the Tictanic.xlsx file on our course web site.



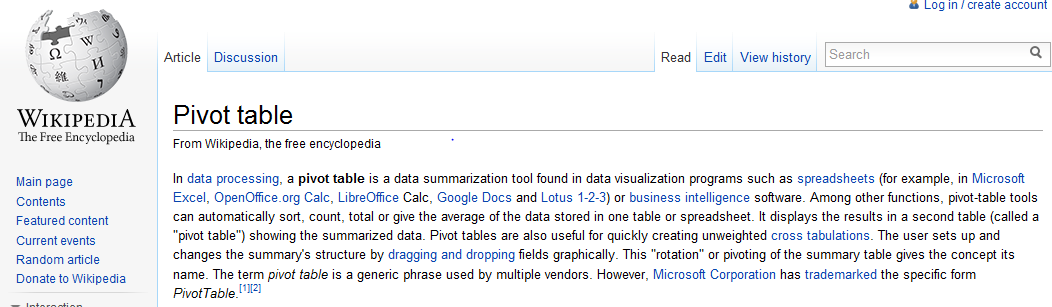
Images of R.M.S. Titanic floorplan.



Investigative Task: Investigate the relationship between the survival rate and passenger class. Possible questions to investigate:

* Is it true that 1st class passengers were more likely to have survived than 3rd class passengers?
* Where 2nd class passengers more or less likely to have survived than 3rd class passengers?
* What percent of the passengers survived?
* What proportion of the passengers were 1st class passengers?

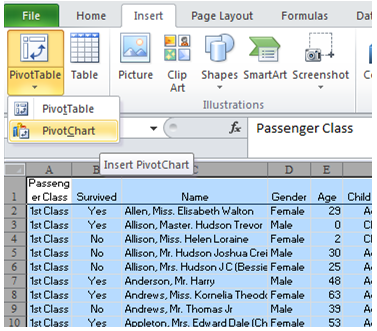
The Excel feature that will allow us to summarize this type of data is called a PivotTable.



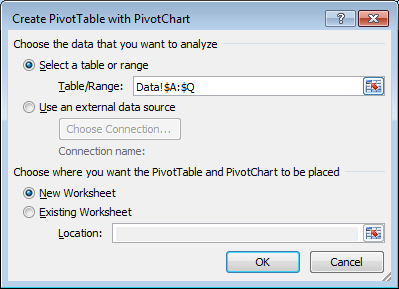
PivotTables are one of the most under-utilized tools in Excel when it comes to analyzing data. PivotTables are powerful and permit us to slice the data is various ways which is important when understanding relationships between variables.

Getting a PivotTable in Excel

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

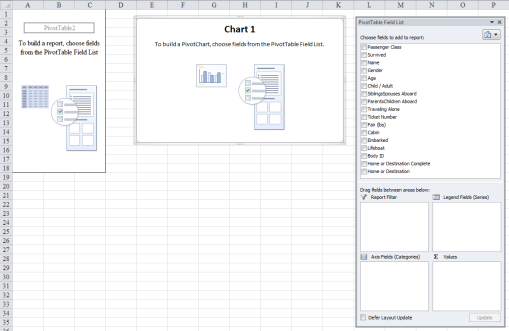


The Create PivotTable window will appear. If you are you have previously highlighted the data that you want to summarize, it will appear the Select a table or range box. The data to be summarized should be included in this box.

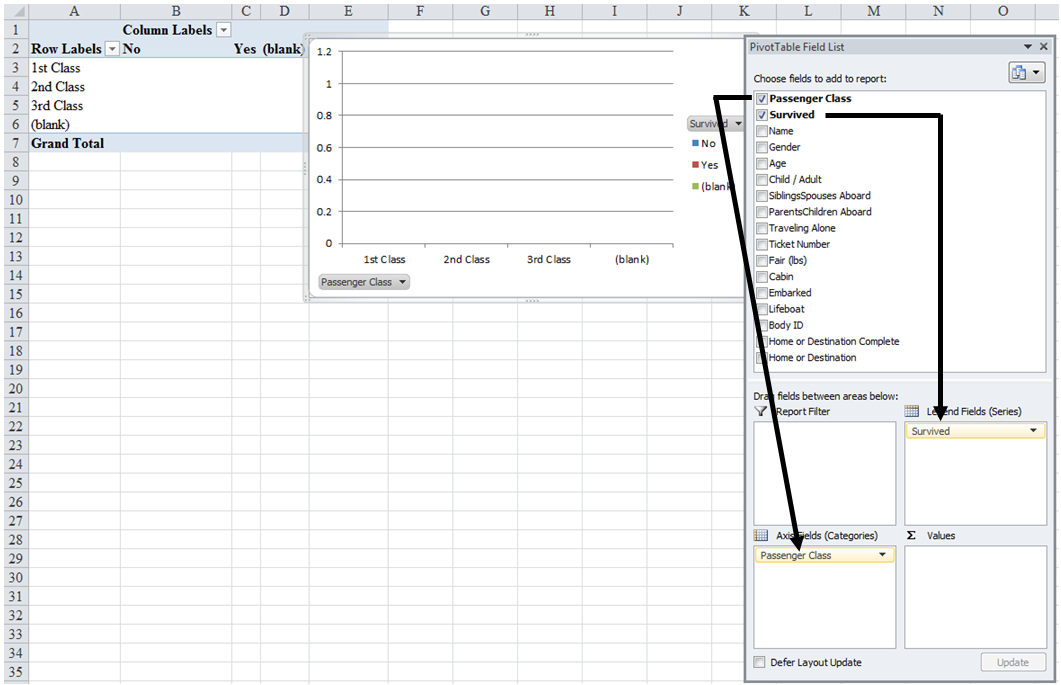


The selection near the bottom specifies where the PivotTable / PivotChart will be placed. I typically place the PivotTable on a New Worksheet. Click OK.

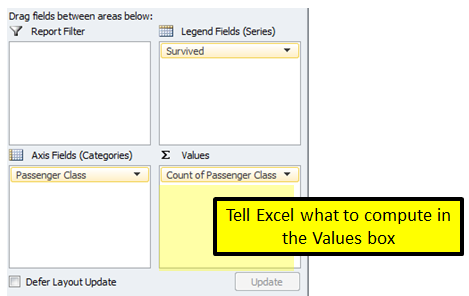
The following is returned by Excel. This template is the basic template for a PivotTable and/or PivotChart.



To create a PivotTable and/or PivotChart, simply select the variables you’d like to summarize and place them in either a row or column. The predictor variable is typically placed in the rows of the table and the response variables in the columns of the table.

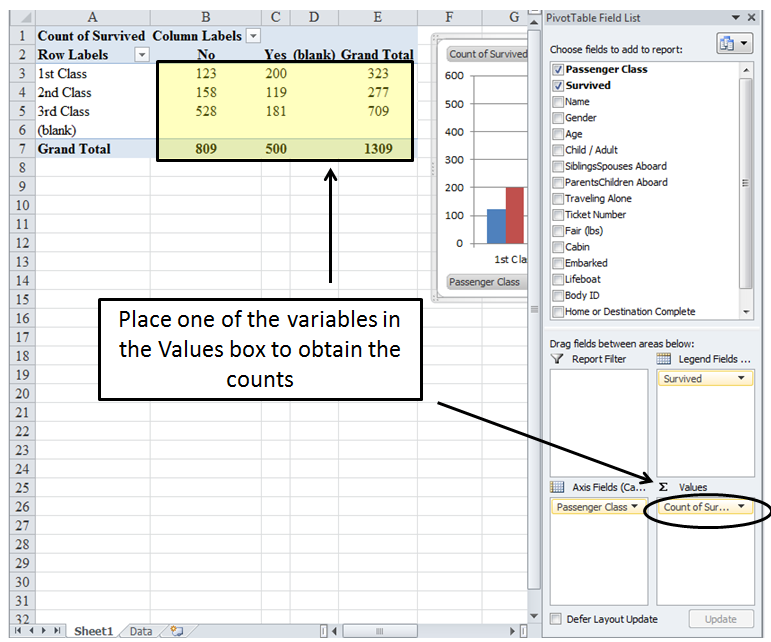


Notice, that nothing has been summarized in the table. To obtain summarizes, drop one of the variables under investigation into the Values box in the lower right-hand corner.



PivotTable setup in Excel

* The **response** variable (i.e. Survived) in the columns
* The **predictor** variable (i.e. Passenger Class) in the rows
* The **Count** of Survived in the Values box



Use your PivotTable to obtain the approprite counts for the following table. Statisticians refer to this type of summary as a **contingency table** or cross-tab table (cross-tab is short for cross-tabular).

|  |  |  |  |
| --- | --- | --- | --- |
| Passenger  Class | Survived | | Total |
| No | Yes |
| 1st Class |  |  | 323 |
| 2nd Class |  |  | 277 |
| 3rd Class |  |  | 709 |
| Total | 809 | 500 | 1309 |

Questions

1. How many first class passengers were on the Titanic? How many 3rd class passengers?
2. How many passengers survived? Did a majority of the passengers survive?
3. Your friend make the following statement, “More 3rd class passengers survived than 2nd class passengers (181 vs. 119).” This statement is technically true, but why is such a statement misleading? Explain.

When making comparisions between groups, we must take into consideration that the groups may be of different sizes. The most common approach to alleviate this concern, is to compute **row percentages**.

Questions

1. Consider only the 3rd class passengers. How many 3rd class passengers were there? How many survived?
2. What proportion of the 3rd class passengers survived?
3. Consider next the 2st class passengers. What proportion of the 2nd class passengers survived?
4. Compare the proportion of 3rd class passengers that survived to the 2nd class passengers that survived? How different are these proportions?
5. Your friend disagrees with how you computed the survival rate above and instead uses the total number of passengers in the denominator.

Survival Rate for 2nd class passengers: 119/1309 ≈ 9%

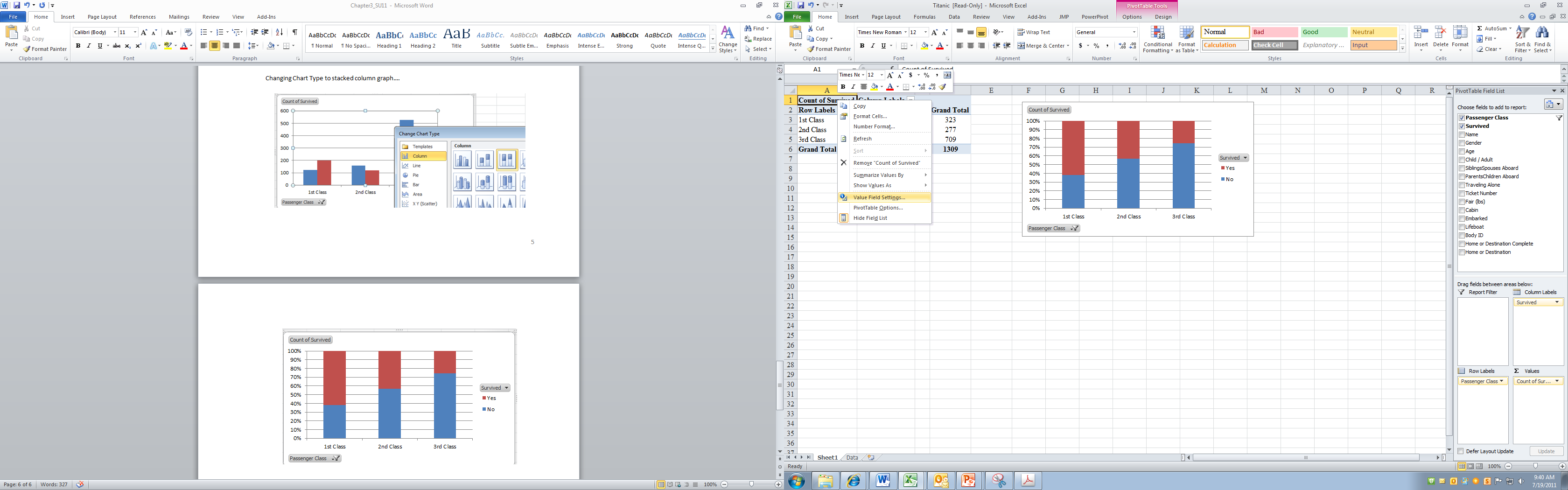
Survival Rate for 3rd class passengers: 181/1309 ≈ 14%

Compare these proportions.

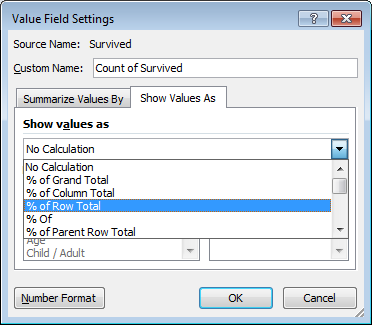
1. Do the comparisons made in Question 7 agree with those in Question 8? If they differ, which comparisons are more meaningful? Discuss.

Getting Row Percentages in Excel

Excel can create the necessary row percentages so that fair comparisons can be made. To obtain the row percentage, right click on the Count of Survived cell in your PivotTable. Select Value Field Settings…



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



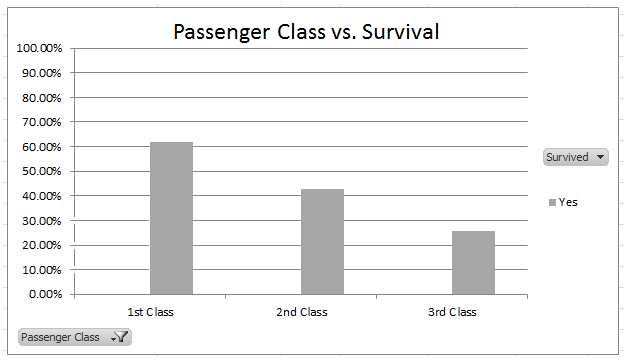
Use your PivotTable output to obtain the missing row percentages in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| Passenger  Class | Survived | | Total |
| No | Yes |
| 1st Class | 38.1% | 61.9% | 100% |
| 2nd Class | 57.0% |  | 100% |
| 3rd Class |  | 25.5% | 100% |

Questions

1. Explain how the value 61.9% was comptued in the above table. Show the formula with the actual numbers.
2. 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.
3. Compare the surival rate of 1st class passengers to that of 3rd class passengers. How different are these survival rates?

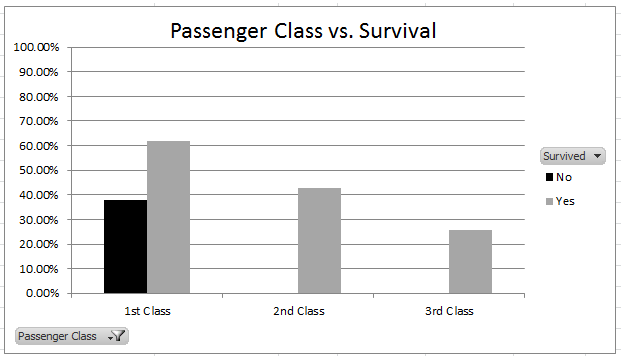
As discussed previously, whenever the totals are different across the groups, row percentages should be used when making comparisons. The following graph shows the survivial rate for each passenger class.



Questions

1. What is the exact height of the 1st Class bar? That is, what values were used in constructing this plot?
2. What pattern(s) do you see in the plot? Discuss.

When a plot of this type is contructed, typically all the row percentages are included – not just the row percentages for those that survived.

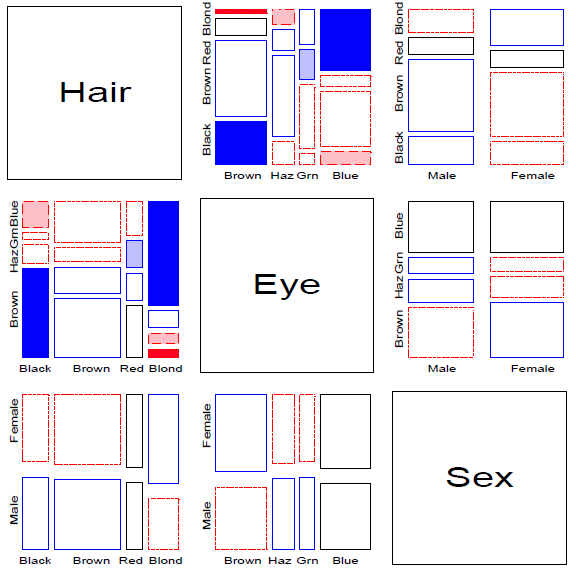


Questions

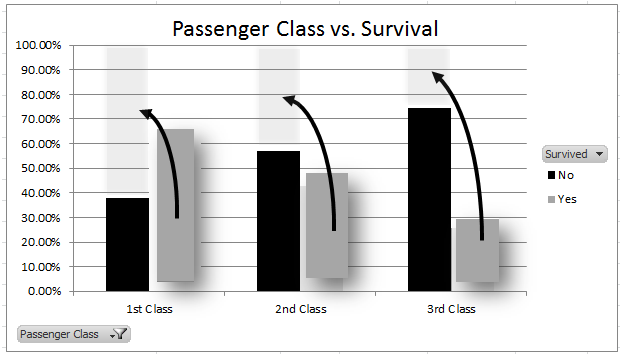
1. On the above plot, sketch the appropriate bar for the proportion that died for the 2nd class and 3rd class passengers. What pattern(s) is/are present when the proportion that died is considered across the passengers classes? Discuss.
2. What is the exact height of the 1st Class bar? That is, what values were used in constructing this plot?

Over time, several people (e.g. Eward Tufte, William Cleveland, Micheal Friendly) have studied how best to display data. Micheal Friendly had devoted much of his work to displaying relationships for categorical data. His webpage is located at <http://www.datavis.ca/>.

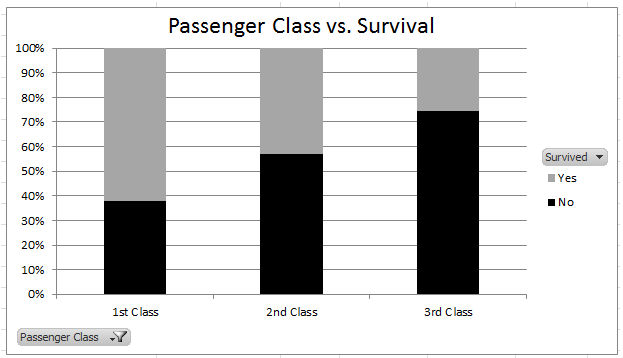
An example of Micheal Friendly’s work is given here.



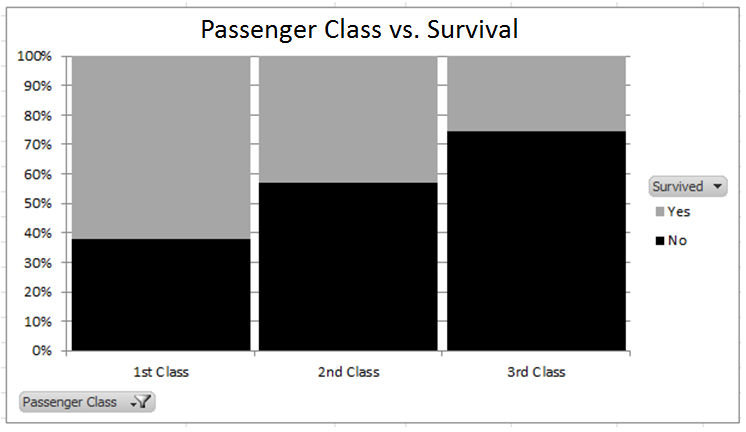
One suggestion to improving our understanding of the relationship between passenger class and survival is to stack the bars for each passenger class instead of using the side-by-side bars.



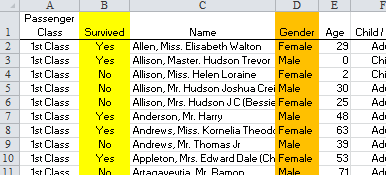
After the bars are stacked, the white space is typically removed between the bars. To remove the white space, select one of the bars, right click and select Format Data Series… In the Format Data Series.. window, reduce the Gap width.



The resulting plot is called a **mosiac plot**.



Example 4.1.2: For this example, consider the relationship between survival rate and gender.



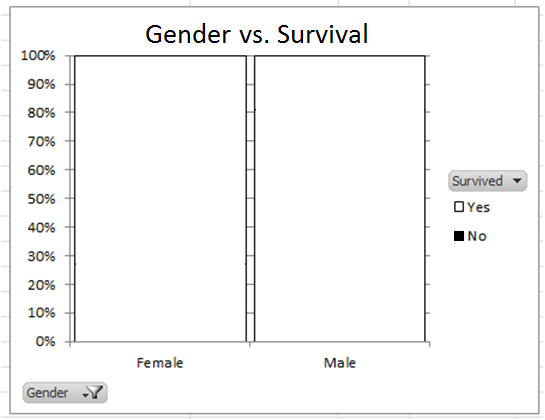
Use the PivotTable feature in Excel to obtain the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| Gender | Survived | | Total |
| No | Yes |
| Female | 127 | 339 | 466 |
| Male | 682 | 161 | 843 |
| Total | 809 | 500 | 1309 |

Obtain the appropriate row percentages using Excel or a calculator.

|  |  |  |  |
| --- | --- | --- | --- |
| Gender | Survived | | Total |
| No | Yes |
| Female |  |  | 100% |
| Male |  |  | 100% |
| Total |  |  | 100% |

Create a mosic plot using the row percentages computed abov e.



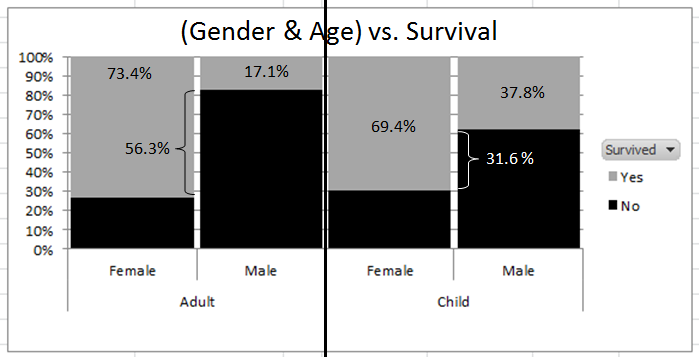
Questions:

1. What was the survival rate of Females on the Titanic?
2. What was the survival rate of Males on the Titanic?
3. How different were the survival rates across Gender? Discuss.

The investigation in Example 4.1.2 considered the relationship between survival rate and gender. From this investigation, we learned that females were more likely to have suvived (see Question 19). The difference in the survival rates between females and males was a little over 50% (i.e. 72.9% - 19.1% = 53.6%).

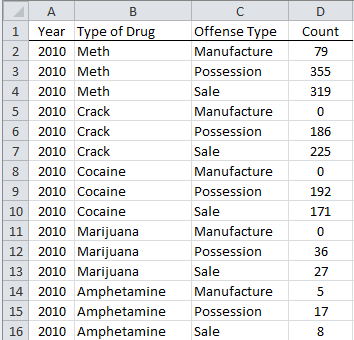
|  |  |  |  |
| --- | --- | --- | --- |
| Survival Rates | | | |
| Adult | | Child | |
| Female | Male | Female | Male |
| 73.4% | 17.1% | 69.4% | 37.8% |
| Difference = 73.4% - 17.1% = 56.3% | | Difference = 69.4% - 37.8% = 31.6% | |
| Notice, the effect of Gender **depends** on Age (i.e. Adult vs. Child) | | | |

The dependence on Age can be seen visually as well.



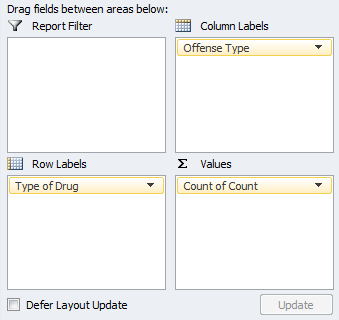
|  |
| --- |
| Definition |
| Interaction between Variables: Interaction is present when the relationship between a predictor variable and response variable is influenced by a second predictor variable. This terminology is reserved for effects or influences amongst the predictor variables.  In the example presented here, the relationship between Gender and Survival is influenced by Age. Gender and Age are the predictor variables in this case and so Gender and Age are said to interact. |

Example 4.1.3: Consider the following data. The relationship between Offense Type and Type of Drug is of interest in this example. The dataset can be downloaded from our course website (i.e. MN Drug Offenses)

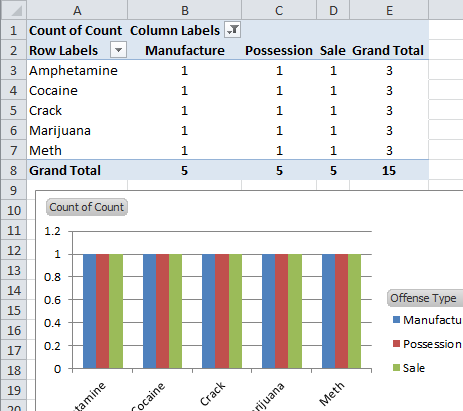


Source: MN Department of Corrections; <http://www.doc.state.mn.us/publications/publications.htm>

Use the PivotTable feature in Excel to obtain a summary of this information. Place Type of Drug as the row, Offense Type as the column, and Count in the Values box.

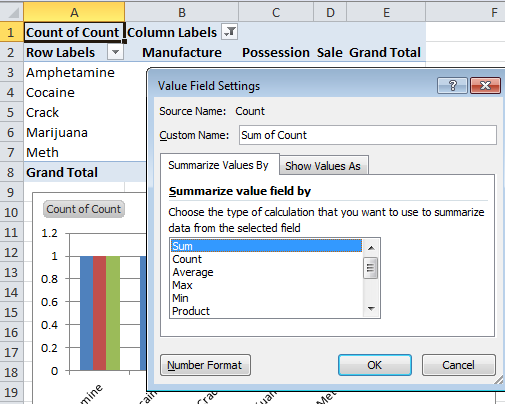


The following table and chart are produced.



This table and graph produced above is an incorrect summary. This summary is incorrect because the PivotTable is simply counting the number of rows in the dataset for each Type of Drug / Offense Type combination. Instead, we should **Sum** the Count variable, not count it.

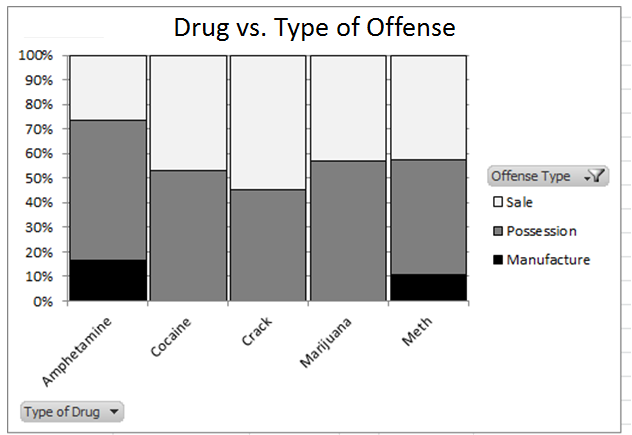
A **sum** of the Count variable can obtained by selecting Sum in the Value Field Settings window.



Below is a partial table of the total number of offenses for each drug. Fill-in the remaining values in this table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Drug | Type of Offense | | | Total |
| Manufacture | Possession | Sale |
| Amphetamine | 5 | 17 | 8 | 30 |
| Cocaine |  | 192 | 171 | 363 |
| Crack |  |  |  | 411 |
| Marijuana | 0 | 36 |  | 63 |
| Meth | 79 |  | 319 | 753 |
| Total | 84 | 786 | 750 | 1620 |
|  | | | | |
| Table of Row Percentages | | | | |
| Drug | Type of Offense | | | Total |
| Manufacture | Possession | Sale |
| Amphetamine | 16.67% | 56.67% | 26.67% | 100% |
| Cocaine | 0.00% | 52.89% | 47.11% | 100% |
| Crack | 0.00% | 45.26% | 54.74% | 100% |
| Marijuana | 0.00% | 57.14% | 42.86% | 100% |
| Meth | 10.49% | 47.14% | 42.36% | 100% |

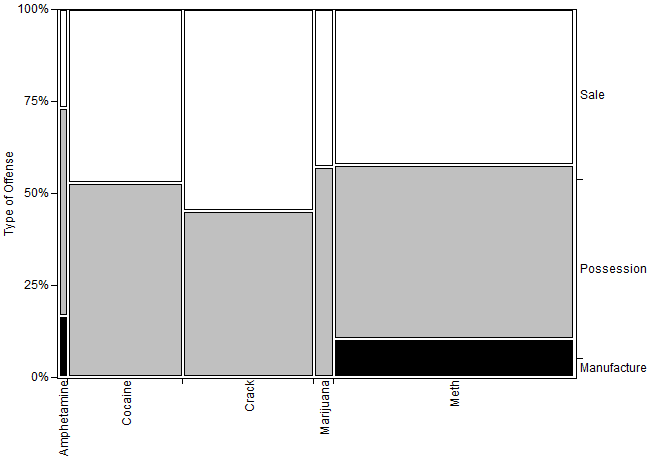
Mosaic plot showing the relationship between Drug and Type of Offense.



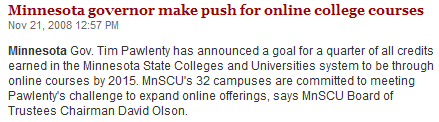
Questions

1. For which drug is Offense Type = Sale most prevalent?
2. For which drug is Manufacture most prevalent?
3. Cocaine, Crack, and Marijuana have no black boxes on this graph. What might this imply about the manufacture of these drugs in Minnesota?
4. Does it appear that police have an appropriate balance in the convictions for the Sale and Possession of drugs in MN? Discuss.

Comment: Some information contained in the data is lost when considering the mosaic plot created in Excel. For example, Meth is the most prevalent drug (753/1620 = 46.5%), but this is not easily seen in the above plot. One approach would be to make the Meth column much wider than the others. In particular, make the Meth column 46.5% of the total width. This is shown here in a modified version of a mosaic plot which is done by the JMP statistical software package.



Example 4.1.4: For many years, people have pushed for more online college courses. Consider the following statement by past Governor Pawlenty. His hope was that 1 in 4 credits earned through the MnSCU system would be from online courses.

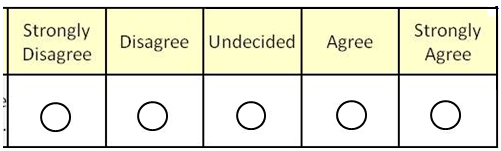


A recent study done was completed on the quality and impact of these types of courses. For our investigation, consider the following key components of quality.

* This online course provided enough opportunities for me to interact with othe students.
* This online course provided enough opportunties for me to interact with the instructor.
* The overall quality of this online course is similar to traditional courses I have taken at this school.

Data was collected from online classes at Winona State and from other MnSCU schools. A 5-point likert scale was used to collect information on each compenent of quality with one exception. A 7-point likert scale was used on the overall quality componenet when collected from other MnSCU schools.

Examples of commonly used 5-point likert scales.



An outdated approach to analyzing this type of data is to convert these categories into numbers. For example, Strongly Disagree would be labeled a “1”, Disagree a “2”, Undecided or Neutral a “3”, Agree a “4”, and Strongly Agree a “5”. I believe this was done because, histroically, data in numeric form was easier to analyze. This is not necessarily the case anymore.



Converting data from one type to another (here from categorical to numeric) may appear conveient, but does not come without consequences. For example, the outcomes from the first key compenent (opporntunties to interact with other students) was converting to a numeric scale and the average was obtained for each instuition type.

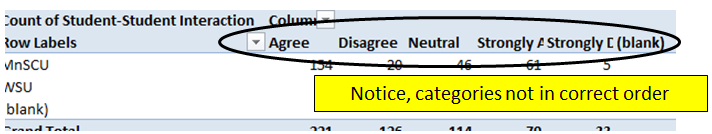
|  |  |
| --- | --- |
| Instuition | Average for Student-to-Student Interaction |
| MnSCU | 3.86 |
| WSU | 2.72 |
| Overall | 3.29 |

The interpretation of these averages often proves to be difficult. We can say, on average, that an individual from MnSCU rated this outcome “*somewhere between Neutral and Agree, but closer to Agree*”. The interpretation for WSU proves to be equally difficult, the average rating for WSU was “*a little short of Neutral, but closer to Neutral than Disagree, but further from Neutral than the MnSCU average was from Agree”.*

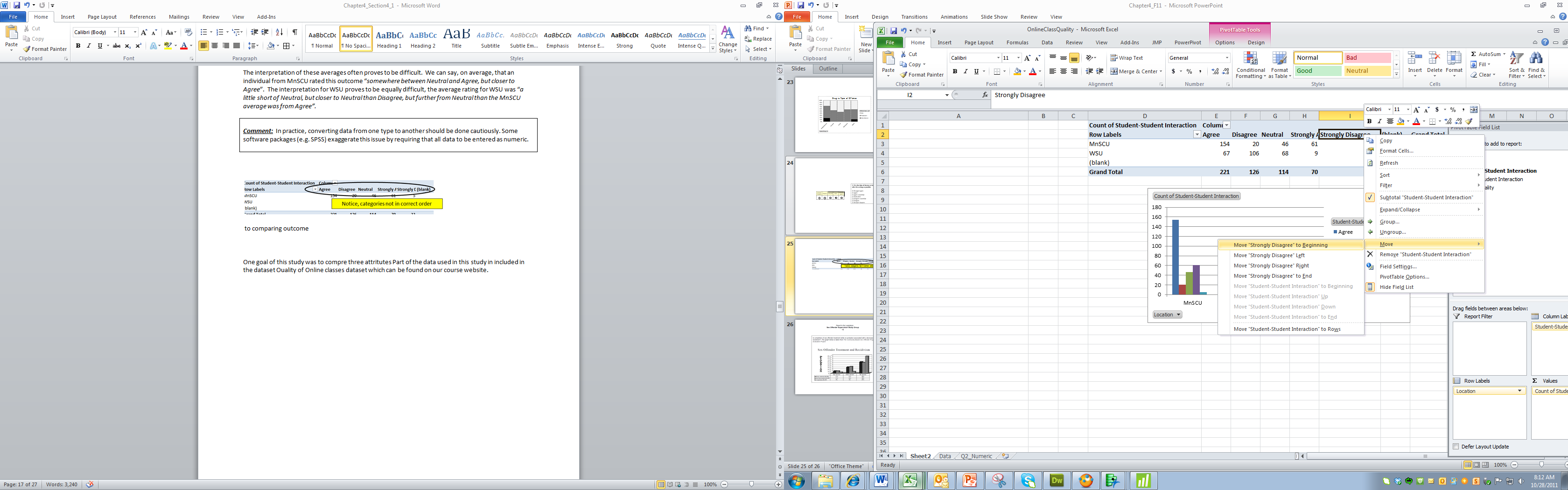
|  |
| --- |
| ***Comment:*** In practice, converting data from one type to another should be done cautiously. Some software packages (e.g. SPSS) exaggerate this issue by requiring that all data to be entered as numeric. |

Obtain the Quality of Online Classes dataset off our course website. Use the PivotTable feature in Excel to investigate the relationship between Location (MnSCU / WSU) to the outcomes for the Student-to-Student Interaction componenet.

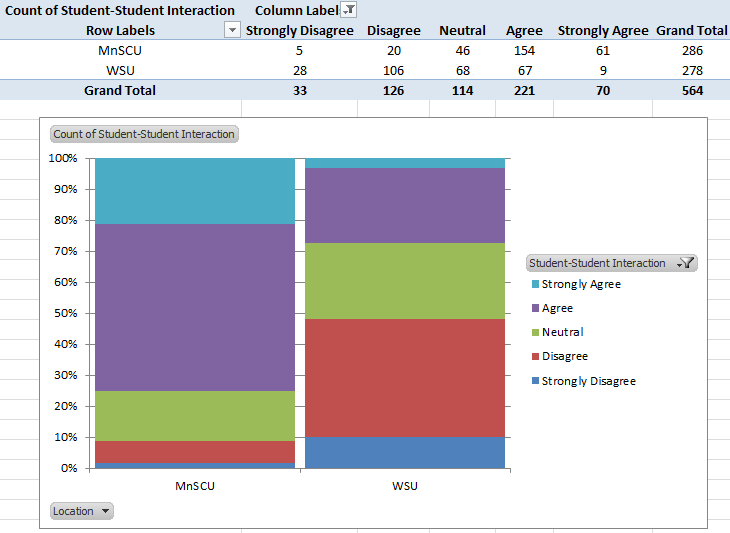
When the PivotTable feautre I used, notice that the outcomes are placed in alphetical order in our table. These categories have a natural ordering and any tables and/or graphs should preserve this ordering.



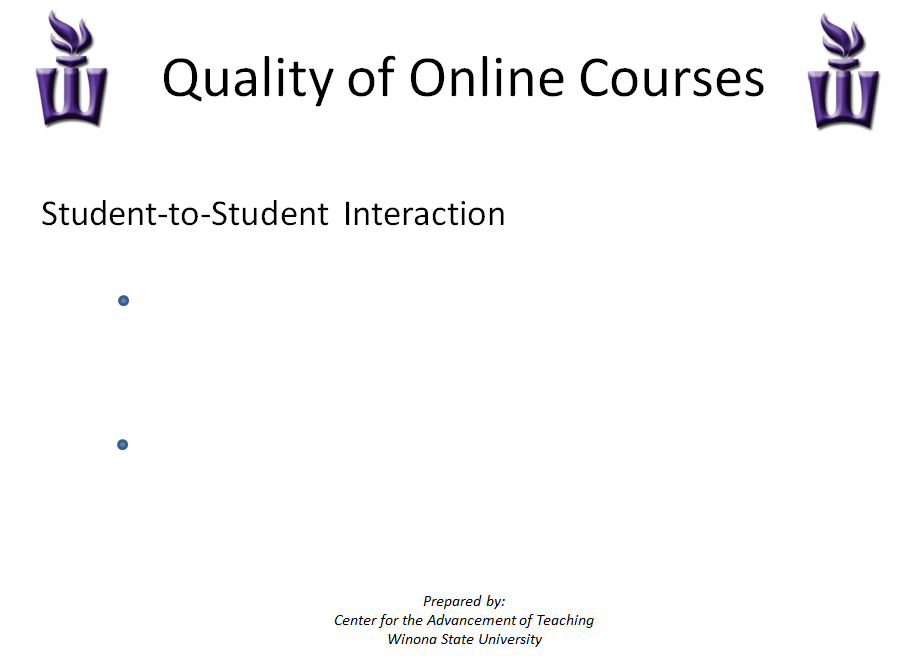
To reagrrange the order, right click on the column you’d like to move and select Move. In the following, Strongly Disagree is moved to the far left.



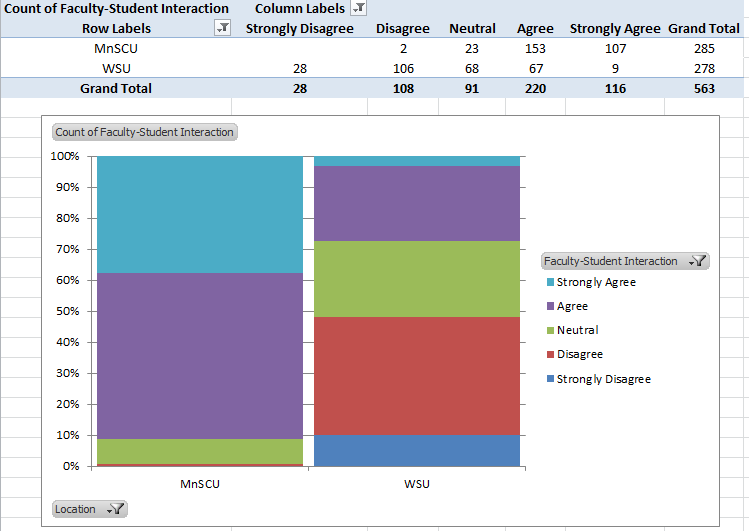
Student-to-Student Interaction Across Instution Type



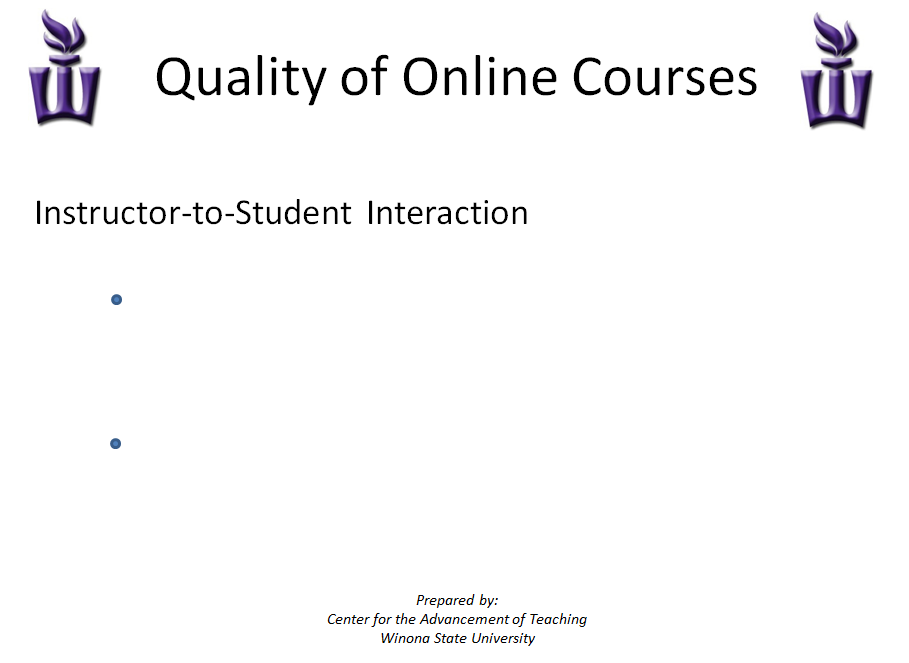
Create two statements regarding the relationship between institution type and opportunities given to students in online courses to interact with other students.



Instructor-to-Student Interaction Across Instution Type



Create two statements regarding the relationship between institution type and opportunities given to students in online courses to interact with their instructor.



Consider the following summary table of the relationship between

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | Instructor-to-Student  Interaction | | | | |
| Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| Student  -to-  Student  Interaction | Strongly Disagree | 28 |  | 1 | 4 |  |
| Disagree |  | 106 | 3 | 11 | 6 |
| Neutral |  | 2 | 78 | 28 | 6 |
| Agree |  |  | 6 | 164 | 47 |
| Strongly Agree |  |  | 2 | 12 | 55 |

Questions

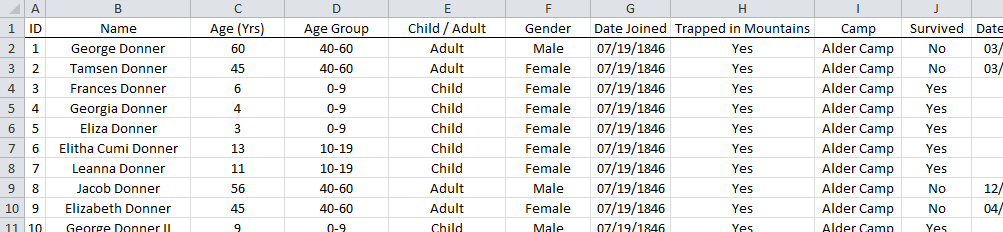
1. Consider the value in the first row and column in this table (i.e. the value 28). Describe what this value represents in our data.
2. What is the connection between all the gray boxes along the diagonal in this table?
3. What do the values above the diagonal represent? What about the value below the diagonal?
4. There were a total of 22 values below the diagonal and 106 above the diagonal. What does this substanital difference represent in the context of this problem? Discuss.
5. Using this data, what advice would you give a new instructor in regards to establishing a strong relationship between students and/or between the instructor-and-students for an online course?

4.2: Methods to Summarize the Relationship

In this section, we will consider methods that can be used to quantify or measure the relationship between two categorical variables. In order to measure an outcome, we typically consider the **risk** or **odds** that a specific outcome of the response variables will occur. Such risks or odds are then contrasted against each other across the levels of the predictor variable in order to understand the relationship between predictor and response variables.

Example 4.2.1: The Donner Party (sometimes called the Donner–Reed Party) was a group of American pioneers who set out for California in a wagon train. Delayed by a series of mishaps, they spent the winter of 1846–47 snowbound in the Sierra Nevada. Information about members of the Donner Party is contained in the DonnerParty.xlsx file on our course website.

A partial listing of the data.



A memorial now stands in this area.

|  |  |
| --- | --- |
| [http://upload.wikimedia.org/wikipedia/commons/thumb/6/66/Map_of_Truckee_Lake_and_Alder_Creek.svg/350px-Map_of_Truckee_Lake_and_Alder_Creek.svg.png](http://en.wikipedia.org/wiki/File:Map_of_Truckee_Lake_and_Alder_Creek.svg) | [Three figures on a tall stone plinth.](http://en.wikipedia.org/wiki/File:Donner_Party_Memorial.jpg)  Statue at Donner Memorial State Park; the top of the 22-foot pedestal indicating how deep the snow was during the winter of 1846–1847. |
| Source: <http://en.wikipedia.org/wiki/Donner_Party> | |

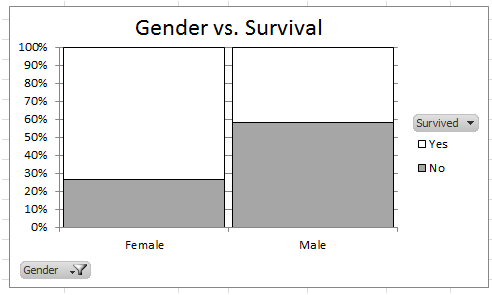
Consider the relationship between Gender and Survival, or more specifically the lack thereof.

Comment: When computing risk (or odds), the *bad* outcome is typically identified as the *risk*.

The following table of counts and the accompanying row percentages were obtained using the PivotTable feature in Excel.

|  |  |
| --- | --- |
| Table of Counts | Table of Row Percentages |
| |  |  |  |  | | --- | --- | --- | --- | | Gender | Survived | | Total | | No | Yes | | Female | 9 | 25 | 34 | | Male | 32 | 23 | 55 | | Total | 41 | 48 | 89 | | |  |  |  |  | | --- | --- | --- | --- | | Gender | Survived | | Total | | No | Yes | | Female | 26.5% | 73.5% | 100% | | Male | 58.2% | 41.8% | 100% | |

Also, a mosaic plot of the row percentages is provided here.



Questions

1. What is the risk of death for Males? Explain how you obtained this value.
2. What is the risk of death for Females?

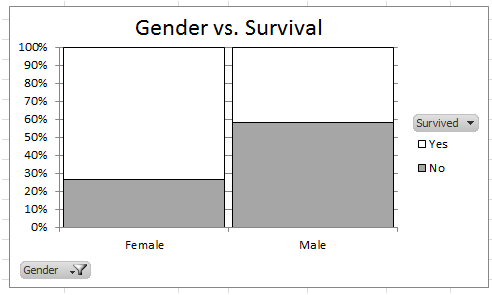
|  |
| --- |
| Measuring Differences using Risk and/or Odds |
| There are two measures that are commonly used to contrast risk values, but only one method to contrast odds.  Comparing Risks: (Should \*not\* be used if row totals are fixed or predetermined by the study design.)   * Difference in Risk, computes the *difference* between the two risks * Risk Ratio, computes the *ratio* between the two risks   Comparing Odds:   * Odds Ratio, computes the *ratio* between two odds |

Difference in Risk

First, consider the calculations for computing the difference in risk.

Risk of Death for Males – Risk of Death for Females = 58.2% - 26.5% = 31.7%

Identify on the mosaic plot the difference in risk.



Interpretation:

Comments:

* If the marginal (i.e. row) total is fixed for any reason, the difference in risks should NOT be computed.
* A difference in risk of zero implies no difference in risk. On the plot below, sketch a scenario in which the difference in risk was zero.
* It is possible for the difference in risk to be less than zero. On the plot below, sketch a scenario in which the difference in risk is below zero.

|  |  |
| --- | --- |
| Difference in Risk = 0 | Difference in Risk < 0 |
|  |  |

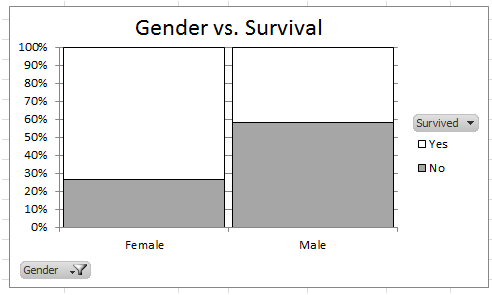
* A small difference in risk is near 0; whereas, a large difference in risk is far away from 0.
* A difference in risk value must be between -100% and 100%.

Risk Ratio

First, consider the calculations for computing the risk ratio.

Risk of Death for Males / Risk of Death for Females = 58.2% / 26.5% = 2.20

Identify on the mosaic plot the risk ratio.



Interpretation:

Comments:

* If the marginal (i.e. row) total is fixed for any reason, a risk ratio should NOT be computed.
* The risk ratio is \*not\* a percent.
* A risk ratio of one implies no difference in risk. On the plot below, sketch a scenario in which the risk ratio is 1.

|  |
| --- |
| Risk Ratio = 1 |
|  |

* The risk ratio is usually computed so that it is larger than 1. For example, we could have computed

Risk of Death for Females / Risk of Death for Males = 26.5% / 58.2% = 0.46

Interpretation: A female was about 0.5 times *more likely* to have died than a male.

* A small risk ratio is near 1; whereas, a large risk ratio is far away from 1.
* A risk ratio cannot go below 0 and has no upper limit.

Odds Ratio

An odds ratio does not use the marginal (i.e. row) total. Thus, an odds ratio can be applied more generally.

|  |  |
| --- | --- |
| Calculate Odds of Death for Females  Odds of Death = 9/25 = 0.36 | Calculate Odds of Death for Males  Odds of Death = 32/23 = 1.39 |
|  |  |
| Odds Ratio = 1.39 / 0.36 = 3.86 | |

Interpretation:

Comments:

* The interpretation of an odds ratio *must* be done in the context of odds.
* An odds ratio of one implies no difference in odds. On the plot below, sketch a scenario in which the risk ratio is 1.

|  |
| --- |
| Odds Ratio = 1 |
|  |

* The risk ratio is usually computed so that it is larger than 1. For example, we could have computed

Odds of Death for Females / Odds of Death for Males = 0.36 / 1.39 = 0.26

Interpretation: The odds of death for female were about 0.26 times *higher* than the   
 odds of death for a male.

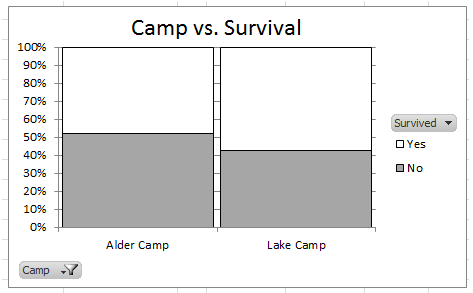
* A small odds ratio is near 1; whereas, a large odds ratio is far away from 1.

Example 4.2.2: Consider the data from the Donner Party again. During the winter of 1846-47, there were two camps that were established in the mountains -- Alder Camp and Lake Camp (i.e. Truckee Lake Camp).

Use the PivotTable feature in Excel to obtain the following counts appropriate to investigate the relationship between Camp and Risk of Death.

|  |  |  |  |
| --- | --- | --- | --- |
| Camp | Survived | | Total |
| No | Yes |
| Alder Camp |  |  | 23 |
| Lake Camp |  |  | 59 |
| Total | 37 | 45 | 82 |

The mosaic plot is given here.



Questions

1. Compute the risk of death for the Alder Camp and for Lake Camp? For which camp was the risk of death higher?
2. Compute the difference in risk? What does this value imply about the differences in risk of death between the two camps? Explain.
3. Compute the risk ratio? What does this value imply about the differences in the risk of death? Explain.

Example 4.2.3 Consider the following information regarding fatalities due to dog bites. According to DogBite.org, there were a total of 88 fatal dog attacks between Jan 1, 2006 and Dec 31, 2008. The following data was given regarding these deaths.

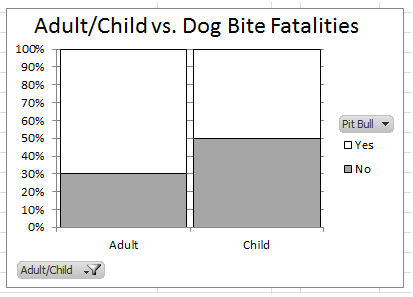
|  |  |
| --- | --- |
|  | |
|  |  |

*Source*: <http://www.dogsbite.org/reports/dogsbite-report-us-dog-bite-fatalities-2006-2008.pdf>

Use the data above to fill in the appropriate counts in the following table. For convience, consider age 20 or less to be child, even though 18 is typically the value used to distingish between child and adult.

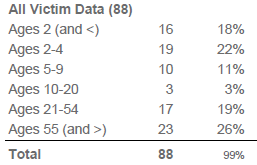
|  |  |  |  |
| --- | --- | --- | --- |
| Age Group | Pit Bull Cause | | Total |
| No | Yes |
| Adult 20 and over |  |  |  |
| Child 20 and under |  |  |  |
| Total |  |  | 88 |

The following mosaic plot was created from this data.



Questions

1. First, consider the data on all victims.

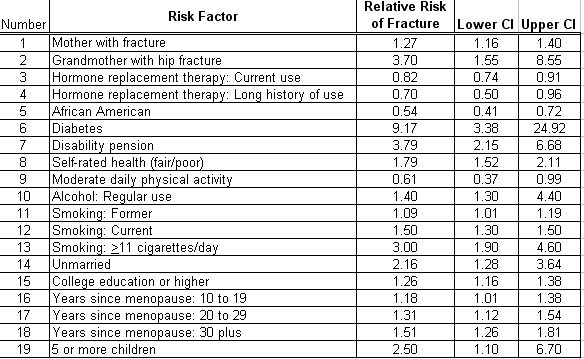


What proportion of dog bite fatalities were children (i.e. age of 20 or under)? Is this value higher, lower, or close to what you expected? Discuss.

1. Next, consider the mosaic plot. The proportion of the adults killed by pit bulls is about 70% or so. How was this value computed?
2. What proportion of the children fatalities were from pit bulls?
3. It is true that adults are more likely to be killed by pit bulls than children? Use a risk ratio to express the increase (or decrease) in risk of a fatality due to a pit bull for adults compared to children.
4. Consider the percentages reported in the summary table above? Would these percentages be considered row percentages? In particular, can these be used directly to construct our mosaic plot? Discuss.
5. Suppose you work for the Winona County Humane Society. How might information from this data and analysis be used to develop a flyer for new owners of pit bulls? Discuss.

Example 4.2.3 Consider the following study in which various risk factors were being considered as a method for screening for postmenopausal osteoporosis.  
  
Source: <http://www.ahrq.gov/clinic/3rduspstf/osteoporosis/osteosumm1.htm>

The risk factors under consideration have been numbered from 1 – 19 in the following table.



As an example, consider the 1st Risk Fracture (Mother with Fracture). The reported relative risk was computed as follows.



Questions

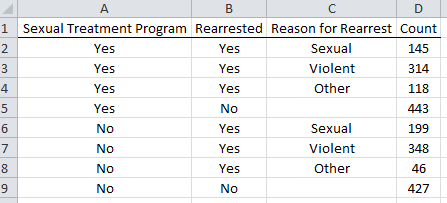
1. Consider the relative risk for Diabetes at 9.17. Using everyday language, explain what this value means?
2. Consider the relative risk for being an African American at 0.54. Using everyday language, explain what this value means?
3. Look at Risk Factors #3 and #4. What can be said about the long term use of hormone replacement therapy in relation to bone fractures in females? Discuss.
4. Look at Risk Factors #11, #12, and #13. What can be said about the effects of smoking in relation to bone fractures in females? Discuss.
5. The risk factors listed above were those found to be statistically important. In particular, notice that none of the confidence intervals capture 1.0. Why is it the case that none of the confidence intervals contain 1.0? What would it mean if the confidence interval did contain 1.0? Explain.

Example 4.2.4 Consider the following data from the MN Department of Corrections web site. The investigation here is centered around whether or not sexual treatment programs work. Consider the following statement in their report.

“To evaluate the effectiveness of sex offender treatment programming, the DOC (Department of Corrections) examined the recidivism outcomes among 2,040 sex offenders released from prison between 1990 and 2003. Recidivism data were collected on 2,040 offenders through 2006. ….. Untreated and treated offenders were matched on commonly known risk factors, and multivariate statistical analyses were performed to control for other factors besides the treatment that may have an impact on recidivism. These measures were used to ensure that ‘apples were compared to apples’.”

Source: “The Impact of Prison-Based Treatment on Sex Offender Recidivism: Evidence from Minnesota”, *Research in Brief*, Minnesota Department of Corrections, March 2010.

The following is some of the data provided in their report.



The data is summarized here.

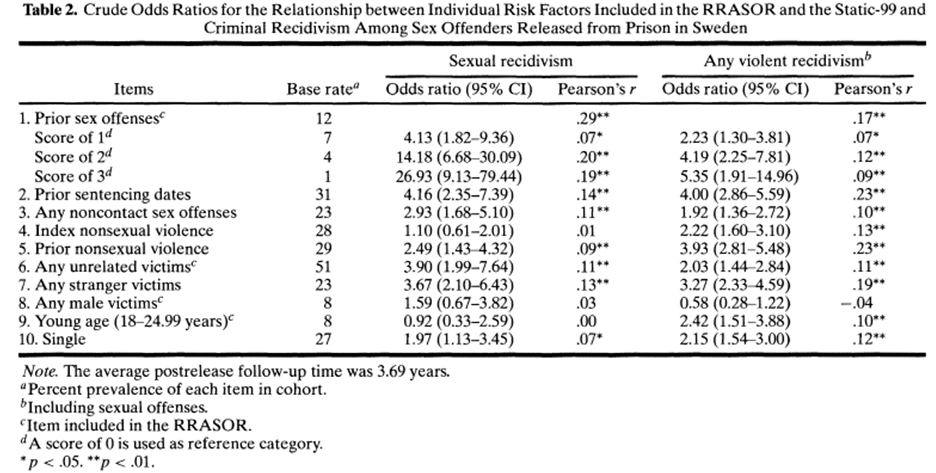
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sexual Treatment Program | Number Not  Rearrested | Number Rearrested  With Reason | | | Total |
| Sexual | Violent | Other |
| Yes | 443 | 145 | 314 | 118 | 1020 |
| No | 427 | 199 | 348 | 46 | 1020 |

Questions

1. Why can’t computations using risk (i.e. difference in risk or risk ratio) be used for this example? Explain.
2. Compute the odds of a rearrest for a person not in a sexual treatment program.
3. Compute the odds of a rearrest for a person who was in a sexual treatment program.
4. Compute the appropriate odds ratio that would allow us to compare the odds of rearrest for

those that did not go through a sexual treatment program to those that did. What are the practical implications of this value? Discuss.

Example 4.2.5 Consider the following table of odd ratios for similar data from a study on sexual recidivism on individuals released from prison in Sweden



*Source*: Sjöstedt, G. and Långström, N. (2001) “Actuarial Assessment of Sex Offender Recidivism Risk: A Cross-Validation of the RRASOR and the Static-99 in Sweden.” *Law and Human Behavior*, Vol. 25, No. 6, pp. 629-645

Questions

1. What are the most important factors that influence the sexual recidivism in this study? What are the least?

The asterisk denotes the statistical significance of each item. Notice, that for each item that lacks statistical significance (i.e. does not have an asterisk), the 95% confidence interval contain 1. This should be the case. Why?

4.3: Statistical Test for 2x2 Tables

In this section, we consider methods that allow us to *test* whether or not data tends to statistically support a research question. The test presented in this section applies to contingency tables that have two rows and two columns. In particular, the predictor variable has two levels and the response variable has two levels. The name of this test is **Fisher’s Exact Test**. A more general test is presented in the next section that allows for more than two levels for the predictor and/or response variable.

Example 4.3.1: Consider once again the Dog Bite data presented in Section 4.2. Recall, this study included the collection of all dog bite fatalities from 2006 through 2008.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data comparing Age to Pit Bull Cause   |  |  |  |  | | --- | --- | --- | --- | | Age Group | Pit Bull Cause | | Total | | No | Yes | | Adult 20 and over | 12 | 28 | 40 | | Child 20 and under | 24 | 24 | 48 | | Total | 36 | 52 | 88 | | Mosaic plot |

Research Question: Can we say, in general, that Adults are more likely to be killed by pit bulls than children?

If this data is a representative sample of all dog bite fatalities, then a statistical test will permit us to generalize (i.e. infer) the outcomes to all dog bite fatalities. This test may best be understood by considering the mosaic plot. The research question asks whether or not, over repeated samples, the “Yes” box for Adults remains larger than the “Yes” box for a Child. If this is the case, then a more general statement can be made regarding the increase in likelihood that an adult will be killed by a pit bull compared to a child.

A p-value will again be used to measure the extremeness of our observed data *under the assumption that age has no effect on whether or not a pit bull was the cause of death.*  This assumption is akin to a 50/50 scenario as was done in Chapter 2.

In the following, identify possible outcome that would be considered more extreme than the observed data (while maintaining the row totals and column totals). When determining extreme, it is important keep the row totals and column totals the same as the original data because the concept of extreme changes if the row and column totals change.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Identify the most extreme table | Most extreme table   |  |  |  |  | | --- | --- | --- | --- | | Age Group | Pit Bull Cause | | Total | | No | Yes | | Adult 20 and over |  |  | 40 | | Child 20 and under |  |  | 48 | | Total | 36 | 52 | 88 | | Mosaic plot |
| Much more extreme than data | Much more extreme than data   |  |  |  |  | | --- | --- | --- | --- | | Age Group | Pit Bull Cause | | Total | | No | Yes | | Adult 20 and over |  |  | 40 | | Child 20 and under |  |  | 48 | | Total | 36 | 52 | 88 | | Mosaic plot |
| Only slightly more extreme than data | Only slightly more extreme than data   |  |  |  |  | | --- | --- | --- | --- | | Age Group | Pit Bull Cause | | Total | | No | Yes | | Adult 20 and over |  |  | 40 | | Child 20 and under |  |  | 48 | | Total | 36 | 52 | 88 | | Mosaic plot |

Determining the Expected Number for each Cell  
In the previous tables, we have considered outcomes that would be more extreme than the observed data. Here, we consider what the expected or anticipated cell counts would for adults and children if indeed there was *no effect due to age on whether or not a pit bull was the cause of death.*

Finding the Expected Cell Counts   
(under the assumption of no effect due to age)

|  |  |  |  |
| --- | --- | --- | --- |
| Age Group | Pit Bull Cause | | Total |
| No | Yes |
| Adult 20 and over |  | ? | 40 |
| Child 20 and under |  |  | 48 |
| Total | 36 | 52 | 88 |

Consider the following procedure for finding the expected cell count for Adult:Yes.

Step 1: First, determine the proportion of adult fatalities.

Proportion of Adult Fatalities = 40 / 88 = 0.4545 = 45.45%

Step 2: Next, there were a total of 52 deaths by pit bulls; under the assumption of no age effect,   
 about 45% of these deaths should be adults.

Expected Cell Count for Adult:Yes = 52 \* .4545 = 23.63

The expected number for the other cells follow similarly.

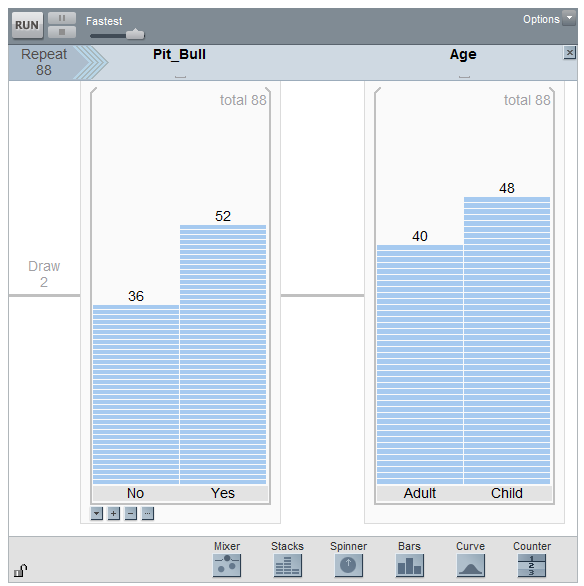
* Expected cell count for Adult:No = 36 \* (40 / 88) = 16.37
* Expected cell count for Child:No = 36 \* (48 / 88) = 19.63
* Expected cell count for Child:Yes = 52 \* ( 48 / 88) = 28.37

The expected cell counts for all cells in our table.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table of Expected Cell Counts (under the assumption of no effect due to age)   |  |  |  |  | | --- | --- | --- | --- | | Age Group | Pit Bull Cause | | Total | | No | Yes | | Adult 20 and over | 16.37 | 23.63 | 40 | | Child 20 and under | 19.63 | 28.37 | 48 | | Total | 36 | 52 | 88 | | Corresponding Mosaic Plot  (under the assumption of no effect due to age) |

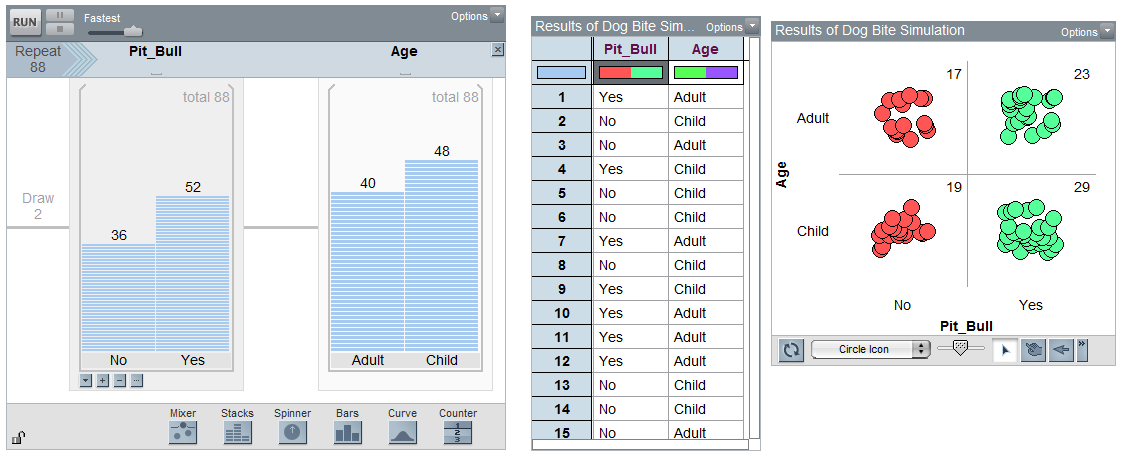
Using Tinkerplots to Obtain the P-Value

The process by which a p-value is obtained is similar to simulations presented earlier. Consider the following setup in Tinkerplots.



* In this mixer, Stacks were used instead of a spinner as was the case with past simulations. The number of observations for each stack is determined by the column and row totals in our data. To specify the number of observations, select Count from the drop-down menu under each stack.
* This simulation requires two stacks each with two outcomes. A second stack is obtained by dragging a new Stack mixer to the right of the original mixer.
* The total number of people in this study was 88; thus, the repeat value is set to 88.
* Finally, the row and column totals cannot change over repeated samples. This would ‘mess up’ what is considered an more extreme. To ensure than row and column totals do not change over repeated samples, select Replacement > Without Replacement from the drop-down menu of each mixer.

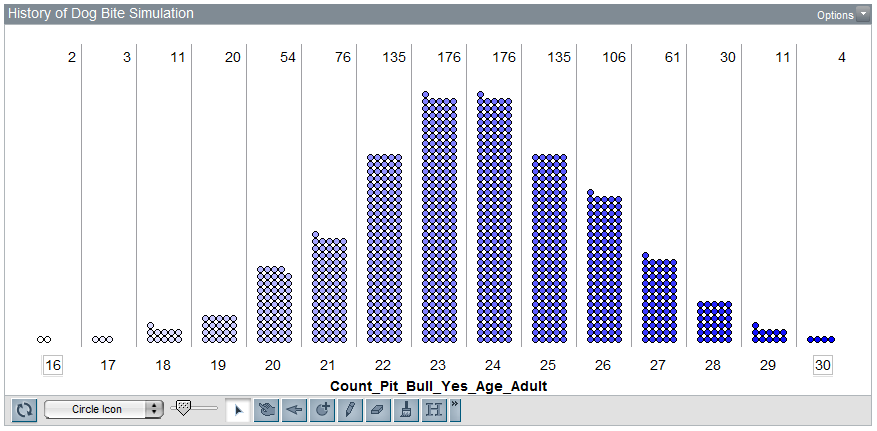
Tinkerplots produces the following outcome for the 1st trial.



Again, it is important to keep the row and column totals the same as the original data for each trial. If this did not happen in your simulation, you likely did not specify that sampling needs to be done without replacement (see note above regarding this issue). For each trial, decide on one statistic to collect for each trial. In the following, Adult: Yes is selected.

|  |  |
| --- | --- |
| Each simulation produces an outcome that maintains row and column totals | Decide on one value to track in each trial. I have selected Adult / Yes here. |

A plot of the outcomes from 1000 trails.



Putting all the pieces together for the Dog Bite case study.

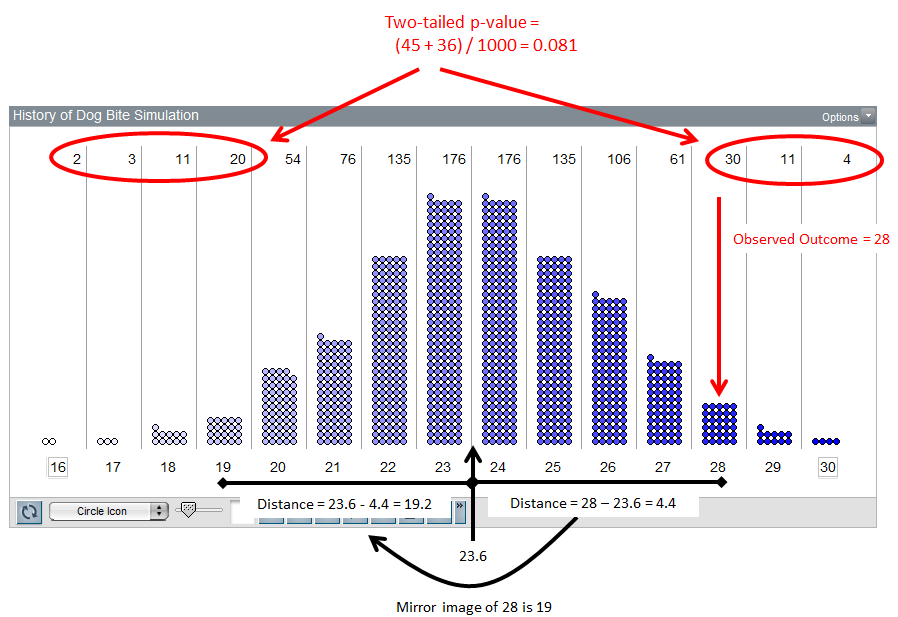
|  |  |
| --- | --- |
| **Dog Bite Case Study** | |
| Research Question | Is there evidence to suggest that Adults are more likely to be killed by pit bulls than Children? |
| Testable Hypothesis | Ho: There is no difference in the chance of being killed by a pit bull between   adults and children  HA: Adults are *more likely* to be killed by pit bulls than Children  Or  Ho: There is no effect due to age in the likelihood of being killed by a pit bull HA: Adults have a higher likelihood of being killed by a pit bull compared to   Children  Or  Ho: Age and Pit Bull Cause are independent of each other HA: Age and Pit Bull Cause are dependent  Or  Ho: Age is not associated with the risk of death by pit bulls HA: An increase is Age is associated with an increase in risk of death by pit bulls  Or ….  Comment: Unfortunately, the language used for this test is not standardized. As a result, several variations exist for how the null and alternative hypotheses are stated. |
| P-Value | p-value = the proportion of outcomes for Adult: Yes that would be as extreme or more extreme than the observed (i.e. 28) that would support the research question    P-Value = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Decision | Decision: If the p-value < 0.05, then data is said to support the research question.   * Data supports research question * Data does not support research question |
| Conclusion | Write a final conclusion using everyday language. |

Comments:

* Suppose the research question was stated as a two-sided test.

Research Question: Is there a difference in risk of death due to pit bulls between adults and   
 children?

The concept of extremeness would then need to include the proportion of dots above 28 and the proportion of dots below 19, which is the mirror image of 28 for this situation.



* If an infinite number of trials were done in Tinkerplots, the resulting reference distribution is called a **hypergeometric** distribution. The computation of the probabilities for a hypergeometric distribution requires the use of combinations.

Concept of Combinations: How many ways can you choose 2 (indistinguishable items) from 3?

|  |  |
| --- | --- |
| Visually | Math Formula for Combinations |

In the context of our dog bite example, the first step is to determine the number of possible ways we can select 12 people from 36. This represents the individuals not killed by pit bulls.

|  |  |  |
| --- | --- | --- |
| One possible way | Another way | In total, 36 choose 12 |

Next, determine the number of ways to select 28 from 52. This combination represents those killed by pit bulls. Putting these two pieces together requires multiplication of these combinations.

|  |  |
| --- | --- |
| 52 choose 28 possible ways for  those killed by pit bulls | Putting the two pieces together  requires multiplication |

Finally, there are 80 choose 40 possible ways to select 40 individuals for the first row. In this case, we do not distinguish between adults and children because probabilities are computed under the assumption that age has no effect.

|  |  |
| --- | --- |
| 80 choose 40 possible ways to obtain 40 individuals in the first row | Putting the pieces together  Probability of observed the 12:28 breakdown in the first row *under the assumption of no effect due to age*  is  Note: This is the similar to the estimated probability obtained from our Tinkerplots simulation (i.e. 30 / 1000 = 0.03). |

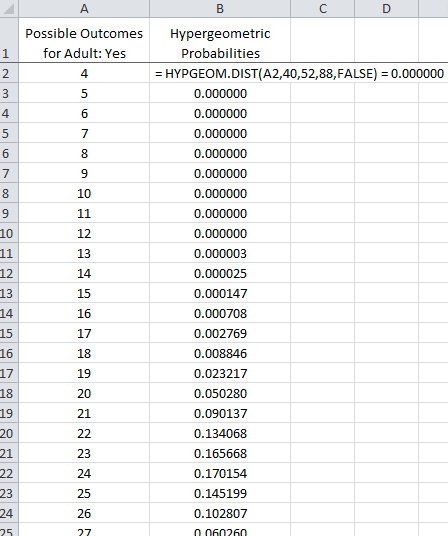
Using Excel to Obtain the P-Value

In Excel, a list of possible values for the cell Adult:Yes must be obtained first.

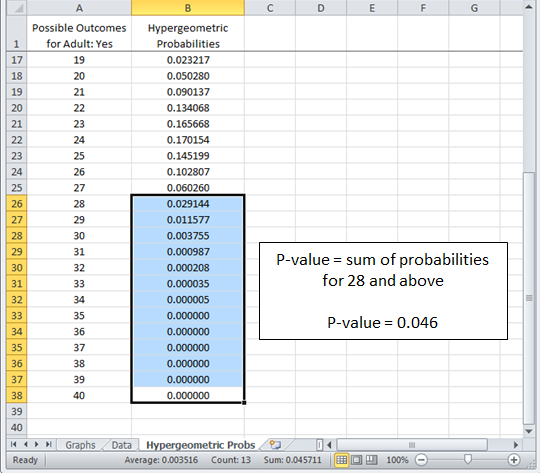
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Smallest possible value for Adult:Yes   |  |  |  |  | | --- | --- | --- | --- | | Age Group | Pit Bull Cause | | Total | | No | Yes | | Adult 20 and over | 36 | 4 | 40 | | Child 20 and under | 0 | 48 | 48 | | Total | 36 | 52 | 88 | | Largest possible value for Adult:Yes   |  |  |  |  | | --- | --- | --- | --- | | Age Group | Pit Bull Cause | | Total | | No | Yes | | Adult 20 and over | 0 | 40 | 40 | | Child 20 and under | 36 | 12 | 48 | | Total | 36 | 52 | 88 | |

Do the following in Excel to obtain the appropriate hypergeometric probabilities.

* Create a list of possible outcome for the Adult:Yes cell.
* Next, use the HYPGEOM.DIST() function in Excel to obtain the appropriate hypergeometric probabilities.
  + The first argument in the HYPGEOM.DIST() function the possible outcomes. For the first row, cell A2 contains the possible outcome. The function will be copied down for all other outcomes.
  + The second argument is the number of total adults, i.e. 40.
  + The third argument is the number of successes, i.e. 52 deaths.
  + The fourth argument is the total number of people in the study, i.e. 88.
  + The last argument is FALSE. The produces the probability for each of the possible outcomes.



To obtaint the p-value for the Dog Bite case study, simply sum the individal probabilities for 28 (i.e. the observed outcome in our study) and above.



Recall, the one-tailed p-value obtained from the Tinkerplots simulation was 0.045. The exact p-value (obtained from the hypergeometric distribution) is 0.046. These values are very similar as would be expected. In practice, statisticians use the hypergeometic distribution directly and skip the simulation-based approach as was done in Tinkerplots. The simulation-based approach is mimicking the hypergeometric distribution, and using the hypergeometic distribution provides an exact p-value for our test.

Example 4.3.2: Consider again the MN Department of Correction investigation into how well sexual treatment programs work in preventing recidivism.

|  |  |  |  |
| --- | --- | --- | --- |
| Sexual Treatment Program | Number Not  Rearrested | Number  Rearrested | Total |
| Yes | 443 | 577 | 1020 |
| No | 427 | 593 | 1020 |

Research Question: Do sexual treatment programs have *any* impact on the likelihood of recidivism (i.e a rearrest)?

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Smallest possible value for Yes: Rearrested   |  |  |  |  | | --- | --- | --- | --- | | Sexual Treatment Program | Not  Rearrested | Rearrested | Total | | Yes | 870 | 150 | 1020 | | No | 0 | 1020 | 1020 | | Total | 870 | 1170 | 2040 | | Largest possible value for Yes: Rearrested   |  |  |  |  | | --- | --- | --- | --- | | Sexual Treatment Program | Not  Rearrested | Rearrested | Total | | Yes | 0 | 1020 | 1020 | | No | 870 | 150 | 1020 | | Total | 870 | 1170 | 2040 | |

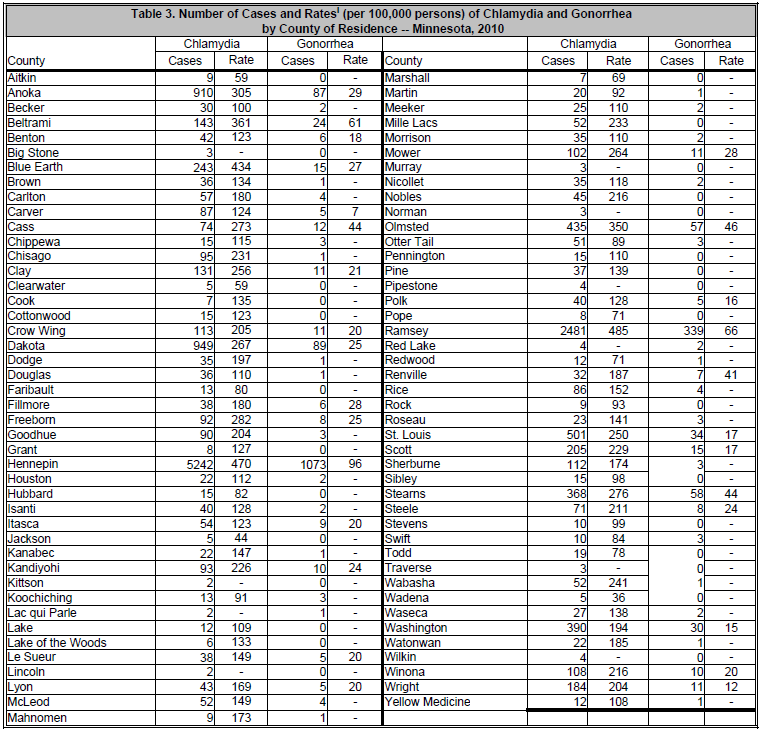
|  |  |
| --- | --- |
| **MN Department of Correction -- Sexual Recidivism Case Study** | |
| Research Question | Do sexual treatment programs have any impact on the likelihood of recidivism? |
| Testable Hypothesis | Ho: There is no effect of the sexual treatment programs on the recidivism rate HA: There is an effect due to the sexual treatment programs on the recidivism   rate |
| Type of Test | Two-tailed test |
| Expected Cell Count | It is necessary to compute the expected cell counts here because we need to identify the mirror image of 577 on our graph.  Compute the expected cell counts for each cell.   |  |  |  |  | | --- | --- | --- | --- | | Treatment Program | Number Not  Rearrested | Number  Rearrested | Total | | Yes |  |  | 1020 | | No |  |  | 1020 | | Total | 870 | 1170 | 2040 | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Observed Data | Observed data from study.   |  |  |  |  | | --- | --- | --- | --- | | Sexual Treatment Program | Number Not  Rearrested | Number  Rearrested | Total | | Yes | 443 | 577 | 1020 | | No | 427 | 593 | 1020 |   In Excel, the Yes:Rearrested value will be used to compute the two-tailed p-value. |
| P-Value | The observed data is 577 and the expected cell count is 585. What is the mirror image of 577 for this problem?  Mirror image value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Now, the p-value is the sum of the hypergeometic probabilities as extreme as or more extreme than the observed data (i.e. 577). Thus, we will need to sum the probabilities from 577 and below and from 593 and higher.    P-Value = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Comment: Individuals have created online applets to allow us to more easily compute the p-value for Fisher’s Exact Test. Also, statistical software packages will compute this p-value automatically. For example, the following output is from JMP, a commonly used statistical software package here at Winona State. |
| Decision | Decision: If the p-value < 0.05, then data is said to support the research question.   * Data supports research question * Data does not support research question |
| Conclusion | Write a final conclusion using everyday language. |

Example 4.3.3: In this example, consider the fact that most often the purpose of a sexual treatment programs for offenders is to prevent the recidivism of a sex crime. Thus, in this example we will consider those rearrested for a sex crime vs. those arrested for a violent/other crime.

|  |  |
| --- | --- |
| **MN Department of Correction -- Sexual Recidivism Case Study, Part II** | |
| Research Question | Do sexual treatment programs reduce the likelihood of recidivism due to a sex crime? |
| Testable Hypothesis | Ho: There is no effect of the sexual treatment program on the recidivism rate of   sex crimes HA: Those that have gone through the sexual treatment program have a *lower*   risk of recidivism for committing a sex crime |
| Type of Test | One-tailed left |
| Observed Data | Observed data from study.   |  |  |  |  | | --- | --- | --- | --- | | Sexual Treatment Program | Rearrested for a Sex Crime | Rearrested for Violent or Other Crime | Total | | Yes | 145 | 432 | 577 | | No | 199 | 394 | 593 | |
| P-Value | Use Excel to obtain the appropriate p-value for a Fisher’s Exact Test.  P-Value = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Decision | Decision: If the p-value < 0.05, then data is said to support the research question.   * Data supports research question * Data does not support research question |
| Conclusion | Write a final conclusion using everyday language for the Director of Sexual Treatment Programs at Minnestoa Department of Corrections. |

Example 4.3.4: Consider the most recent data provided by the Minnesota Department of Health regarding the reported number of cases of Chlamydia and Gonorrhea. Chlamydia and Gonorrhea are the most prevalent STDs in Minnesota. This data includes the number of cases and the rate (expressed as Number per 100,000 persons) for each county in Minnesota.



Consider the following list of two-tailed research questions that allow us to make comparisons across two different counties.

1. Do differences exist in the rate of chlamydia between Winona County Winona State area) and Stearns County (St. Cloud State area)?
2. Do differences exist in the rate of chlamydia between Winona County and Blue Earth County (Mankato State area)?
3. Do differences exist in the rate of chlamydia between Winona County and Lyon County (Southwest State area)?
4. Do differences exist in the rate of chlamydia between Winona County and Ramsey County (St. Paul area)?

Answer any two of the four research questions given above using a Fisher’s Exact Test.

Research Question # \_\_\_\_\_\_\_\_\_\_\_

Use the following structure when compiling data for each of these research questions.

|  |  |  |  |
| --- | --- | --- | --- |
| County | Type of STD | | Total |
| Chlamyida | Gonorrhea |
| Winona |  |  |  |
|  |  |  |  |
| Total |  |  |  |

Obtain the two-tailed p-value using Excel or any other resource.

P-Value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Write a final conclusion using everyday language.

Research Question # \_\_\_\_\_\_\_\_\_\_\_

Use the following structure when compiling data for each of these research questions.

|  |  |  |  |
| --- | --- | --- | --- |
| County | Type of STD | | Total |
| Chlamyida | Gonorrhea |
| Winona |  |  |  |
|  |  |  |  |
| Total |  |  |  |

Obtain the two-tailed p-value using Excel or any other resource.

P-Value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Write a final conclusion using everyday language.

4.4: Statistical Test for Any Size Table

The Fisher Exact Test presented in the previous chapter was applicable to only tables with two rows and two columns (i.e. 2x2 tables). The test presented here is a generalization of this test and can be used on any size table. The test presented here is commonly known as a **Chi-Square Test of Independence**.

Example 4.4.1: Consider once again the MN Department of Correction’s case study on sexual treatment programs.

|  |  |
| --- | --- |
| **MN Department of Correction -- Sexual Recidivism Case Study** | |
| Research Question | Do sexual treatment programs have any impact on the likelihood of recidivism? |
| Testable Hypothesis | Ho: There is no effect of the sexual treatment programs on the recidivism rate HA: There is an effect due to the sexual treatment programs on the recidivism   rate |
| Type of Test | Two-tailed test |
| Expected Cell Counts | Recall, the *expected* cell counts for each cell.   * Yes:Not Rearrested = 870 \* (1020 / 2040) = 435 * Yes:Rearrested = 1170 \* (1020 / 2040) = 585 * No:Not Rearrested = 870 \* (1020 / 2040) = 435 * No:Rearrested = 1170 \* (1020 / 2040) = 585  |  |  |  |  | | --- | --- | --- | --- | | Treatment Program | Number Not  Rearrested | Number  Rearrested | Total | | Yes | 435 | 585 | 1020 | | No | 435 | 585 | 1020 | | Total | 870 | 1170 | 2040 | |
| Observed Data | The *observed* data from the study.   |  |  |  |  | | --- | --- | --- | --- | | Sexual Treatment Program | Number Not  Rearrested | Number  Rearrested | Total | | Yes | 443 | 577 | 1020 | | No | 427 | 593 | 1020 | |

|  |  |
| --- | --- |
| P-Value | The Chi-Square Test of Independence can be done directly in Excel using the **CHITEST()** function in Excel. Similar to Chapter3, the observed and expected cell counts are required by this CHITEST() function.    P-Value = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Comment: The Chi-Square Test of Independence is always a two-tailed test. A one-tailed version does not exist. The p-value obtained here is similar to the two-tailed p-value from the Fisher’s Exact test that was done in the previous section. |
| Decision | Decision: If the p-value < 0.05, then data is said to support the research question.   * Data supports research question * Data does not support research question |
| Conclusion | Write a final conclusion using everyday language. |

A Chi-Square Test of Independence does a cell-by-cell comparison between the observed counts and the expected counts. This formula for this test is exactly the same as what was done in Chapter 3.

Computing the test statistic for the previous example is shown here.

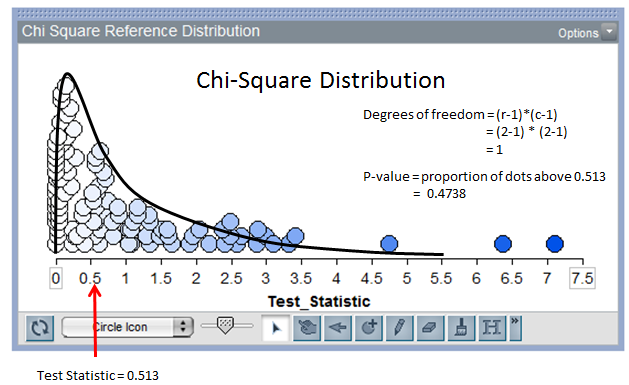
|  |  |  |
| --- | --- | --- |
| Sexual Treatment Program | Number Not  Rearrested | Number  Rearrested |
| Yes | = 0.1471 | = 0.1094 |
| No | =0.1471 | =0.1094 |

The test statistic is the sum of all the cell, test Statistic = 0.1471 + 0.1094 + 0.1471 + 0.1094 = 0.513

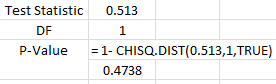
Test Statistic = 0.513

Under the null hypothesis, the appropriate reference distribution is the Chi-Square distribution.

* The Chi-Square distribution is indexed by its degrees of freedom. The degrees of freedom for this test is equal to (# rows in table – 1) \* (# of columns in table – 1).
* The p-value is simply the proportion of dots as extreme or more extreme than the observed test statistic.

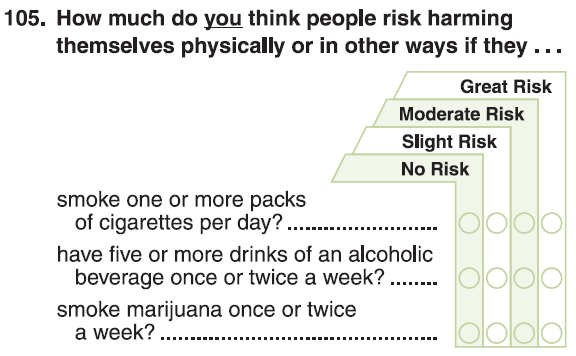


* Given the test statistic and the degrees of freedom (i.e. df), the p-value can be computed using the CHISQ.DIST() function in Excel.



The test statistic of 0.513 is not extreme in this distribution; thus, we have lack statistical evidence to support the research question (p-value = 0.4738).

Example 4.4.2: Consider once again the data from the Minnesota Study Survey regarding the opinions of marijuana use for 12th grade students in Fillmore County.



The following data was obtained from the Minnesota Department of Education website for Fillmore County.



*Source*: Minnesota Department of Education; <http://education.state.mn.us/MDE/Learning_Support/Safe_and_Healthy_Learners/Minnesota_Student_Survey/index.html>

Consider the following investigation comparing Males to Females for Grade 12.

|  |  |
| --- | --- |
| **Grade 12 Analysis: Comparing Genders** | |
| Research Question | Do differences exist between Genders in the opinions regarding marijuana use for 12th graders in Fillmore County? |
| Testable Hypothesis | Write the appropriate null and alternative hypothesis for your investigation.  Ho: Opinions regarding marijuana are independent of Gender  HA: Opinions regarding marijuana depend on Gender |
| Type of Test | Two-tailed test |
| Observed Data | The *observed* data from the study.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Gender | No Risk | Slight Risk | Moderate Risk | Great Risk | Total | | Female | 3 | 8 | 15 | 35 | 61 | | Male | 9 | 8 | 14 | 28 | 59 | | Total | 12 | 16 | 29 | 63 | 120 | |
| Expected Cell Counts | Recall, the *expected* cell counts for each cell.   * Female: No Risk = 12 \* (61/120) = 6.10 * Female: Slight Risk = 16 \* (61/120) = 8.13 * Female: Moderate Risk = 29 \* (61/120) = 14.74 * Female: Great Risk = 63 \* (61/120) = 32.03 * Male: No Risk = 12 \* (59/120) = 5.90 * Male: Slight Risk = 16 \* (59/120) = 7.87 * Male: Moderate Risk = 29 \* (59/120) = 14.26 * Male: Great Risk = 63 \* (59/120) = 30.98  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Gender | No Risk | Slight Risk | Moderate Risk | Great Risk | Total | | Female | 6.10 | 8.13 | 14.74 | 2.03 | 61 | | Male | 5.90 | 7.87 | 14.26 | 30.98 | 59 | | Total | 12 | 16 | 29 | 63 | 120 | |
| P-Value | Getting the Chi-Square Test of Independence in Excel.    P-Value = \_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Decision | Make the appropirate statistical decision.  Decision: If the p-value < 0.05, then data is said to support the research question.   * Data supports research question * Data does not support research question |
| Conclusion | Write a final conclusion that would be appropriate for the DARE administrator from Fillmore County. |

Example 4.4.3: Next, we will investigate the possible association between opinions of marijuana use and grade level. Certainly, young children have a belief, typically a strong belief, that drugs are “bad”. The question here centers around the possible shift in opinions from Grade 9 and Grade 12 students from Fillmore County.

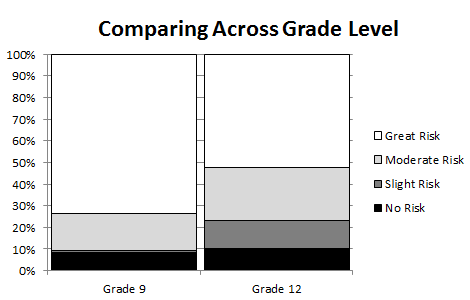


Consider the following investigation which compares across grade levels 9 and 12.

|  |  |
| --- | --- |
| **Comparing Grade 9 to Grade 12** | |
| Research Question | Are there differences in the opinions regarding marijuana use between Grade 9 and Grade 12 students from Fillmore County? |
| Testable Hypothesis | Write the appropriate null and alternative hypothesis for your investigation.  Ho: Opinions regarding marijuana are independent of Age (Grade 9 & 12)  HA: Opinions regarding marijuana depend on Age (Grade 9 & 12) |
| Type of Test | Two-tailed test |
| Observed Data | The *observed* data from the study.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Grade Level | No Risk | Slight Risk | Moderate Risk | Great Risk | Total | | Grade 9 | 14 | 2 | 29 | 127 | 172 | | Grade 12 | 12 | 16 | 29 | 63 | 120 | | Total | 26 | 18 | 58 | 190 | 292 | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Expected Cell Counts | Compute the *expected* cell counts for the missing cells.  Show your work for at least one of these computations.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Grade Level | No Risk | Slight Risk | Moderate Risk | Great Risk | Total | | Grade 9 | 15.32 |  | 34.16 |  | 172 | | Grade 12 |  | 7.40 |  | 23.84 | 120 | | Total | 26 | 18 | 58 | 190 | 292 | |
| P-Value | Getting the Chi-Square Test of Independence in Excel.    P-Value = 2.3753E-05 = 0.000023753 |
| Decision | Make the appropirate statistical decision.  Decision: If the p-value < 0.05, then data is said to support the research question.   * Data supports research question * Data does not support research question |
| Conclusion | Write a final conclusion that would be appropriate for the DARE administrator from Fillmore County. |

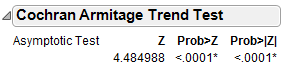
Consider the mosaic plot which can be used to compare across grade levels



Questions:

1. How does this plot agree (or disagree) with the conclusion of the above test?
2. There appears to be a trend present in this plot. Carefully explain these trends. What do these trends imply about opinions regarding marijuana use across Grade 9 vs. Grade 12 students from Fillmore County?

Comment: A **Cochran Armitage Trend Test** can be used to verify whether or not the trend see in the above mosaic plot is likely to appear on repeated sampling from the population. Search Wikipedia for additional information regarding this test. Excel cannot do this test, but the following output was obtained in JMP.



The one-sided p-value for the Cochran Armitage Trend test is less than 0.0001 (i.e. Prob > Z value).

1. What is the practical interpretation of this p-value? Write a statement regarding this test that would be useful for a DARE officer in Fillmore County.

Example 4.4.4 Consider the following data presented in the HIV / AIDS Surveillance Report 2013 on the Minnesota Department of Health web site. The information presented here is comparing the mode of exposure for diagnosis across race/ethnicity.

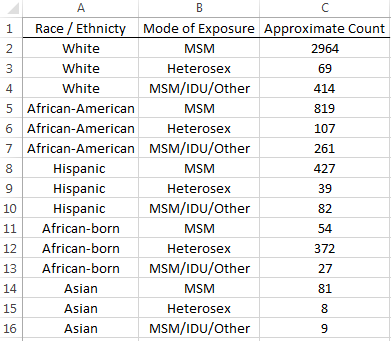
Source: Minnesota Department Health (Link: <http://www.health.state.mn.us/divs/idepc/diseases/hiv/hivsurvrpts.html> )

|  |  |
| --- | --- |
|  | How to read the White Males graph   * A total of 3447 white males are living with HIV or AIDS in MN * 86% (about 2960) were exposed due to MSM (men having sex with men) * 2% were exposed through heterosexual contact * About 11% were exposed from IDU (Injecting Drug Use) or MSM/IDU combination |
|  |  |
|  |  |

Questions

1. Most of the above pie charts contain a large portion of red. What does this mean?
2. The African-born pie chart is strikingly different. How might the HIV / AIDS crisis in several African countries contribute to this difference in trends?
3. What would be the likely outcome of a Chi-Square Test of Independence for testing the relationship between mode of exposure and race? Explain your answer.

I obtained the counts based on the percentages presented above for White, African-American, Hispanic, African-born, and Asian and put them into the following table.



A mosaic plot that shows the trends across race.

|  |
| --- |
| Comparing Mode of Exposure Across Race |

Questions

1. Is it easier or more difficult to make comparisons across race using a mosaic plot or individual pie charts? Explain.
2. Does this plot suggest that a relationship exists between the mode of exposure and race? If so, briefly discuss this relationship.
3. The p-value for the Chi-Square Test of Independence is *very* small -- much smaller than 0.05. What is the practical interpretation of this p-value?
4. Suppose you are in charge of creating a pamphlet for the Minnesota Department of Health. Our investigation here suggests that an HIV/AIDS pamphlet appropriate for a clinic in certain area (say Winona) may need to be changed for a clinic in a different area (say a Somalia Community Health Clinic). Carefully explain why the p-value from our analysis supports this conclusion. Also, what is *gained* by using a p-value to help us make this decision? Why should we even look at a p-value in an investigation like this? Discuss.
5. Is it unethical to investigate data in the manner in which we have here? That is, should HIV / AIDS surveillance reports include information regarding race? Discuss.

Example 4.4.5 Consider a study done by Timonen, et al. (2001) that investigating the risk of suicide and it possible relationship with income level. This study was done in Finland because it has one of the highest death rates from suicide and treatment methods are similar across income levels (i.e. most of the population is treated similarly in public hospitals, regardless of socioeconomic status). Consider the following data from their study.

|  |  |
| --- | --- |
| Violent\* vs. Non Violent Across Occupations    \*Violent Includes: Hanging, drowning, shooting, wrist cutting, jumping from a height | Reason for Admission into Treatment  Across Occupations |

Source: Timonen, M., Viilo, K., Hakko, H., Väisänen, E., Räsänen, P., and Särkioja, T. (2001). “*Risk of suicide related to income level in mental illness: Psychiatric disorders are more severe among suicide victims of higher occupational level*.” British Medical Journal, Vol. 323 (7306) : 232.

Consider the following research questions.

1. Is the likelihood of a violent suicide dependent on type of occupation?
2. Students tend to be younger and the method of suicide for this age group is likely to be different than other age groups. Remove the student data from the dataset and redo the first research question.

Is the likelihood of a violent suicide dependent on type of occupation (with students removed)?

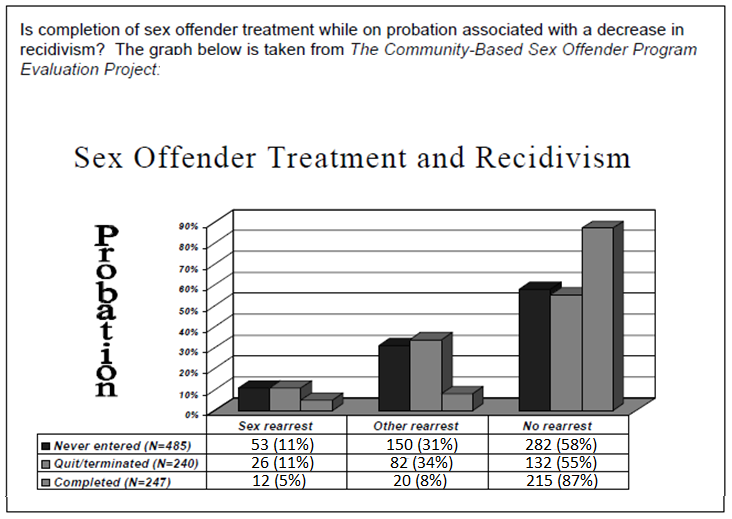
1. Is the likelihood for the reason of admission vary across occupations?

Complete the following for each research question.

* Obtain a p-value from a Chi-Square Test of Independence.
* Answer each research question using your p-value.
* Obtain the appropriate mosaic plot for each research question that can be used to make visual comparisons across the occupations.
* Write two or three statements, appropriate for a mental health clinic, that communicate the practical interpretation of the outcomes from each research question.

Example 4.4.6 The Minnesota Department of Corrections often creates reports for the Minnesota State Legislature to identify the importance of funding programs that are known to improve public safety. The following data was presented in one such report. The following question and supporting data were provided on p17 of their report.

Source: <http://www.doc.state.mn.us/publications/documents/sosupreport.PDF>



Complete the following for your investigation.

* Put this data into Excel and complete a Chi-Square Test of Independence.
* Obtain a mosaic plot for this data.

Questions

1. What is the p-value from your test? In the context of this problem, what is the practical interpretation of this p-value?
2. What is *gained* by running a Chi-Square Test of Independence? Why would somebody actually do this test in an investigation of this type?
3. Notice, their research question is actually one-sided (i.e. interest in a decrease in recidivism), but the Chi-Square Test of Independence is a two-sided test. Explain why a Cochran Armitage Trendtype test (see Example 4.4.3) may actually be a better test for this investigation.