Chapter 3: The Analysis of a Single Categorical Variable across Several Categories

The analyses completed in Chapter 2 were for a single variable with two outcomes. For example, for the Staring Case study, the individual doing the guessing was either correct or incorrect or for the AYP examples, the schools were either making AYP or not making AYP. In this chapter, allow for more than two categories. Extended to more than two categories is easy to simulate in Tinkerplots.

3.1: Understanding Variation in Repeated Samples

Tinkerplots will be used in this section to help us understand how much random variation is acceptable when investigating a single variable with several categories.

Example 3.1.1: The Minneapolis Police Department posts regular updates on crime statistics on their website. I have collected this data for the past two years ( identified as Fiscal Year = Current or Past) on all neighborhoods in Minneapolis. The data and prescient map are given here.

|  |  |
| --- | --- |
| Minneapolis Crime Statistics (see course website) | Precinct Map |

Source: <http://www.minneapolismn.gov/police/crime-statistics/>

The police chief for Precinct #2 has received a complaint from a permanent resident who lives in a neighborhood near the University of Minnesota. This resident has asked for additional patrol to take place in his neighborhood as he believes that crime rates vary over the course of the year.

Research Question: Is there evidence to suggest that crime patterns in the University of Minnesota neighborhood differ over the four seasons of the year?

Crime rates are reported by month, so use the following definitions for the Seasons:

* Fall: September, October, and November
* Winter: December, January, and February
* Spring: March, April, and May
* Summer: June, July, and August

The crimes of Murder, Rape, Robbery, Aggravated Assault, Burglary, Larceny, Auto Theft, and Arson are used in reporting the **Total**. The counts reflect the number of crimes reported and arrests made.

The Minneapolis Police Department reported that a total of 103 crimes for the University of Minnesota neighborhood last year.

|  |  |
| --- | --- |
| **Minneapolis Crime Case Study** | |
| Research Question | Is there evidence to suggest that crime patterns in the University of Minnesota neighborhood differ over the four seasons of the year? |
| Testable Hypothesis | Ho: Crimes are equally dispersed over the four seasons  HA: Crimes are not occurring equally over the four seasons |
| Parameters | The four parameters of interest are defined as follows:  = the probability of a crime occurring in the Fall  = the probability of a crime occurring in the Winter  = the probability of a crime occurring in the Spring  = the probability of a crime occurring in the Summer |
| Rewrite of Hypotheses |  |

The approach taken here to answer the research question is very similar to what we have done previously. We will use Tinkerplots to conduct a simulation *assuming* the crime patterns are occurring equally across the four seasons. We will then check to see if our observed outcomes are outliers against the simulated outcomes. If the observed outcomes are outliers, then we have sufficient statistical evidence to say crimes rates vary of the four seasons.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Season | | | |
| Fall | Winter | Spring | Summer |
| U of MN | 25% | 25% | 25% | 25% |

Questions

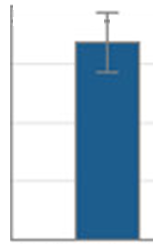
1. What is the number of anticipated or expected outcomes for each season under the assumption that crimes are occurring equally over the four seasons. Carefully, explain how you obtained these values.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Season | | | | Total |
| Fall | Winter | Spring | Summer |
| U of MN |  |  |  |  | 103 |

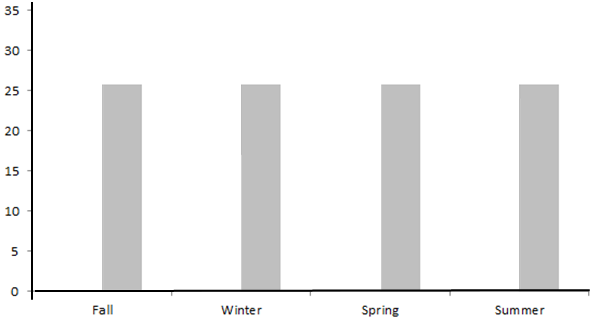
1. One of your sometimes annoying friends asks, “How would I compute the anticipated number if the percentages were not all equal?”. Consider the following percentages. Explain to your friend how to compute the anticipated number for this situation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Season | | | |
| Fall | Winter | Spring | Summer |
| U of MN | 30% | 25% | 25% | 20% |

1. A statistician would argue that we must allow for some slight variations in the crime patterns over the four seasons because we should not expect the numbers to come out exactly at the expected number for each season. Do you agree? Explain.
2. Over repeated samples, slight variations will occur in the crime patterns. The amount of acceptable variations is measured by the margin-of-error and is sometimes displayed on the top of each bar as is shown here.



On the following bar chart, estimate the amount of acceptable random variation for each of the four seasons.



1. In the above plot, is the estimated amount of acceptable variation about the same for each season or different? Explain your rationale.
2. Ask your neighbors what they decided to use as an estimate for the amount of acceptable variation for each season.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Acceptable Amount of Variation (i.e. Margin-of-Error) | | | |
| Fall | Winter | Spring | Summer |
| Neighbor 1 |  |  |  |  |
| Neighbor 2 |  |  |  |  |
| Neighbor 3 |  |  |  |  |

How does your estimate compare to your neighbors for each season? Did your neighbors use the same estimate for each season? Discuss.

Tactile Simulation

In an effort to better understand an appropriate amount of random variation, you and your friend decide to run a simulation. One problem is that you are on a deserted island and all you have (other than fresh water, food, and shelter) is an 8 sided tie and time.

Questions

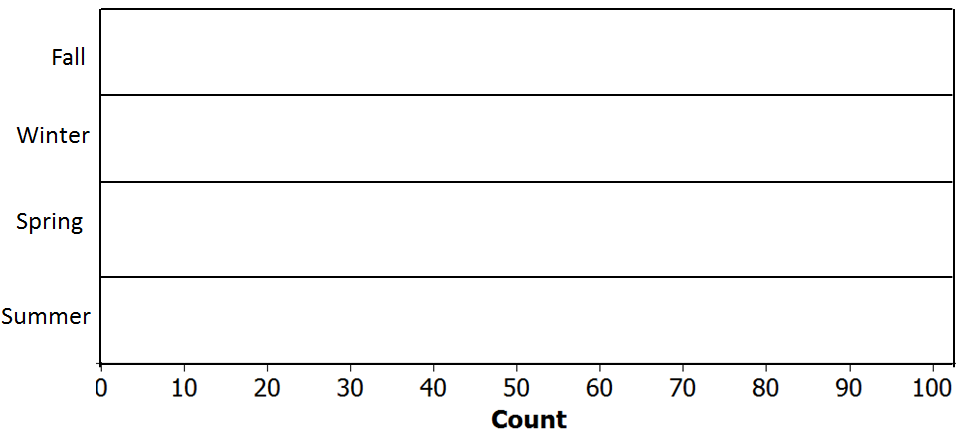
1. Your simulation has four categories instead of two; thus, you need something other than a coin to run your simulation. Now, unfortunately they don’t make a four sided die, but as luck would have it, they do make an eight sided die which can be used to run your simulation. Clearly identify which seasons will be associated with each number of this 8 sided die.

|  |  |
| --- | --- |
| Number on Die | Label for Outcome |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |
| 8 |  |

1. How many times will you need to roll your die to mimic the occurrence of crimes for the University of Minnesota neighborhood over the four seasons? Explain.
2. Your friend makes the following false statement, “A trial consists of four rolls of the die, one for each season.” Why is this statement wrong? What constitutes a trial in this situation? Explain.
3. Consider the following chart which we’ve used in the past to keep track of the outcomes in Tinkerplots for single trial. Give a likely and unlikely set of values for the number of crimes for Fall, Winter, Spring, and Summer from a single trial of your die simulation.

|  |  |
| --- | --- |
| Likely Set of Outcomes | Unlikely Set of Outcomes |

1. Explain how you identified appropriate values for the likely and unlikely situations above.
2. For each trial, you and your friend record the number of crimes that occurred in fall, winter, spring, and summer from your 8 sided die. Plot the anticipated pattern for 10 trials on the number lines below.

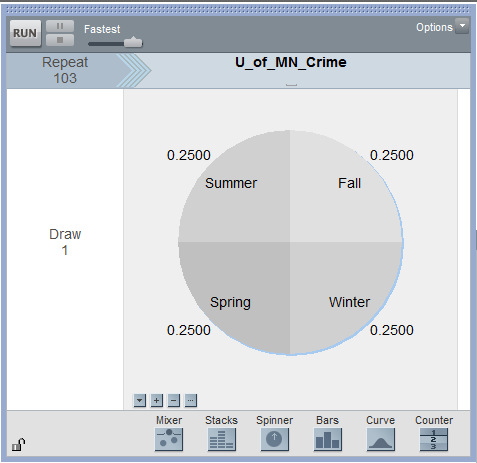


The police chief for Precinct #2 is on vacation and comes upon you and your friend on your deserted island. He decides to rescue you, but under one condition. You have to clearly explain to him what you have been doing with this 8-sided die. Write a brief letter to that describes what you have been doing and the purpose of this simulation. You should address how such a simulation will help answer his original research question.

|  |
| --- |
| Dear Police Chief,    Sincerely,    P>S> Please rescue us! |

Tinkerplots Simulation

TinkerpIots can be used to run a simulation akin to the one performed above. Tinkerplots will allow us to have four categories on the spinner. Create the following spinner in Tinkerplots . Click Run.



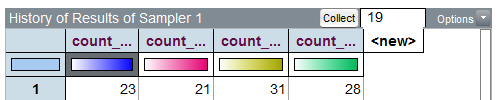
Create a plot similar to one provided below and record the number of crimes for each season.

|  |  |
| --- | --- |
| Plot your outcomes from a single trial | My Outcomes |

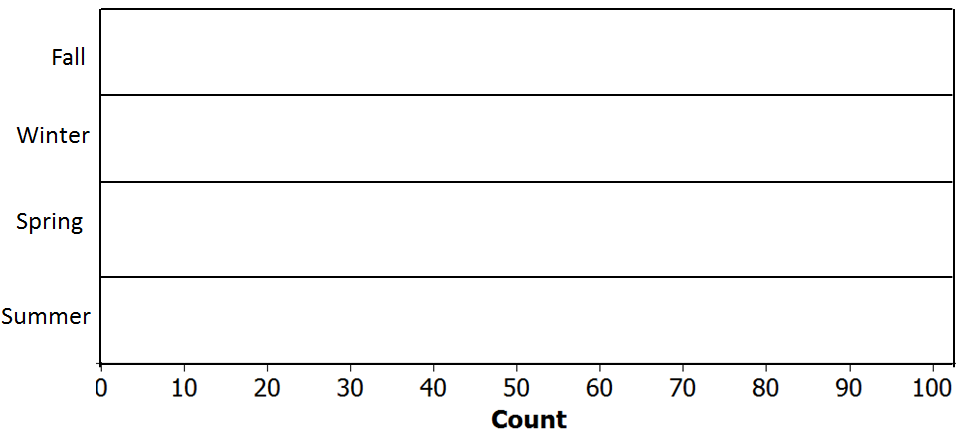
Questions

1. Why is the spinner setup with 25% for each season? Explain.
2. Why is the repeat value set to 103? Where did this number come from? Explain.
3. Did the outcomes from your trial (i.e. the number of crimes for fall, spring, summer, and winter) match mine? Should they match? Explain.

InTinkerplots, obtain the count for the number of crimes for each season. Right click on each count and select Collect Statistic. This will need to be repeated for each season. Once this is done, place 19 in the Collect box so that a total of 20 trials is obtained.



Plot the outcomes for each season. Give a rough sketch of each plot on the number lines below.



On the plot above, identify a reasonable value for a lower cutoff and an upper cutoff for when you start to believe an outcome would be considered an outlier.

* Lower Cutoff: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Upper Cutoff: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Your friend makes the following true statement, “It is reasonable to use the same lower and upper cutoff for each season.” Why is this statement true? Discuss.

Next, consider the actual crime statistics for the University of Minnesota neighborhood for the past year.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Season | | | | Total |
| Fall | Winter | Spring | Summer |
| U of MN | 32 | 17 | 30 | 24 | 103 |

Research Question: Is there evidence to suggest that crime patterns in the University of Minnesota neighborhood differ over the four seasons of the year?

Questions

1. Use the outcomes from your 20 simulation done in Tinkerplots and the observed outcomes to provide a *tentative* answer the research question.

Note: Tentative because a p-value has not been obtained yet.

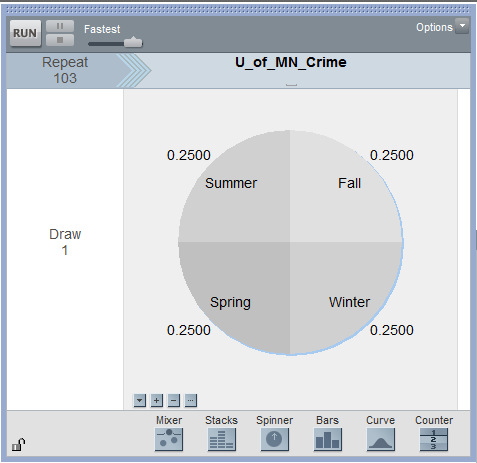
1. Discuss any difficulties when trying to answer this question when four categories are present. Specifically, why is it more difficult to determine whether or not our data is considered an *outlier* in this situation?

3.2: Using Technology to Quantify Variation in Repeated Samples

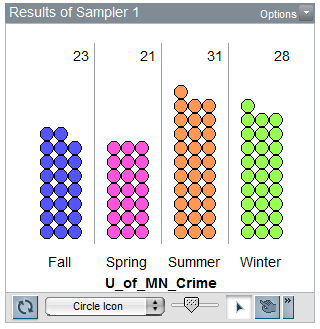
Example 3.2.1: Consider again the Minneapolis Police Department Crime case study. Data for this case study can be found on our course website.

Research Question: Is there evidence to suggest that crime patterns in the University of Minnesota neighborhood differ over the four seasons of the year?

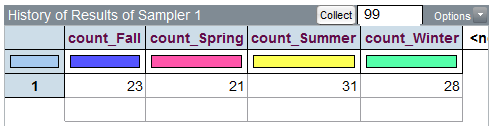
Consider the appropriate spinner setup for this case study in Tinkerplots.



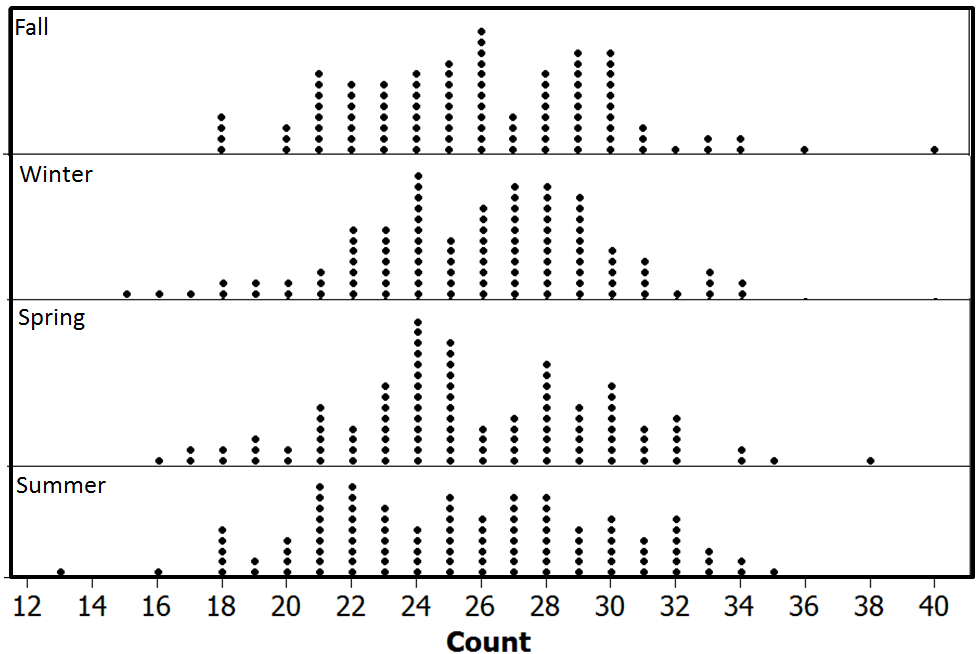
Click Run. Create a plot of the outcomes produced from the first trial.



There are four categories, thus we will need have Tinkerplots Collect Statistic for each category. An additional 99 trials were collected and plotted below.



The outcomes from the 100 trials completed in Tinkerplots are shown here.



Recall, the research question for this case study, “Is there evidence to suggest that crime patterns in the University of Minnesota neighborhood differ over the four seasons of the year?” In order to answer this question, we need to identify whether or not the observed data would be considered an outlier. This needs to be done for each season.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Season | | | | Total |
| Fall | Winter | Spring | Summer |
| U of MN | 32 | 17 | 30 | 24 | 103 |

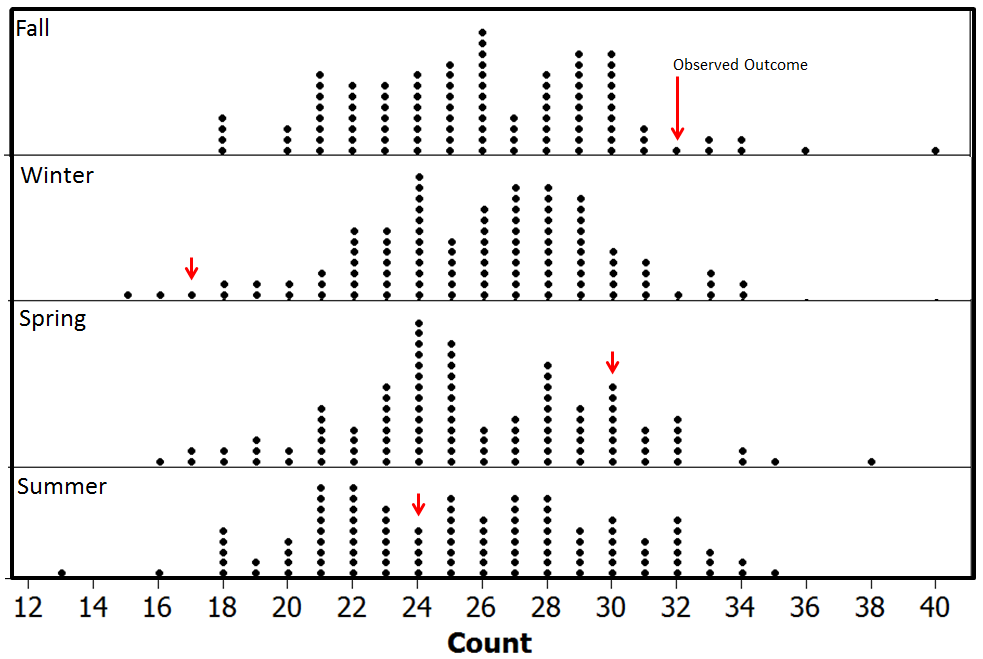
Determine whether or not the outcomes for Winter and Spring would be considered outliers.

|  |  |  |  |
| --- | --- | --- | --- |
| Season | Outlier | | |
| Yes | No | Maybe |
| Fall |  |  | X |
| Winter |  |  |  |
| Spring |  |  |  |
| Summer |  | X |  |

To formalize the concept of an outlier, we will again consider the p-value approach. The definition of a p-value is given here as a reminder.

|  |
| --- |
| P-Value: the probability of observing an outcome as extreme or more extreme than the observed outcome that provides evidence for the research question |

Recall, the research question fro this analysis, “Is there evidence to suggest that crime patterns in the University of Minnesota neighborhood differ over the four seasons of the year?”



Compute the approximate *two-tailed* p-value for each season.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Season | Computing p-value | # Dots  Upper-Side | # Dots  Lower-Side | Total  Dots | Estimated  P-Value |
| Fall | Number of dots more extreme than 32 |  |  |  |  |
| Winter | Number of dots more extreme than 17 |  |  |  |  |
| Spring | Number of dots more extrem than 30 |  |  |  |  |
| Summer | Number of dots more extreme than 24 |  |  |  |  |

Questions

1. Use the p-value computed above to determine whethor or not the data supprots the research question. What is your decision?

Formal 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

1. Discuss any difficulties when trying to answer this question when four categories are present. Specifically, why is it more difficult to determine whether or not our data supports the research question?

Comment: The issue of combining p-values (aka “multiplicity of tests” or simply “multiple comparisons”) to make a single decision has not been universally resolved. Statisticians continue to be required to deal with this issue in practice. The most significant concern when combining p-values is that the familywise (or experiment-wide) error rate is much greater than 0.05, our gold standard for making decisions.

|  |  |  |  |
| --- | --- | --- | --- |
| Season | Estimated  P-Value | Error  Rate | Statistically  Significant |
| Fall | 0.14 | 0.05 | No |
| Winter | 0.03 | 0.05 | Yes |
| Spring | 0.41 | 0.05 | No |
| Summer | 0.78 | 0.05 | No |
| Maximum Error Rate  (across all four comparisons) | | 0.20 |  |

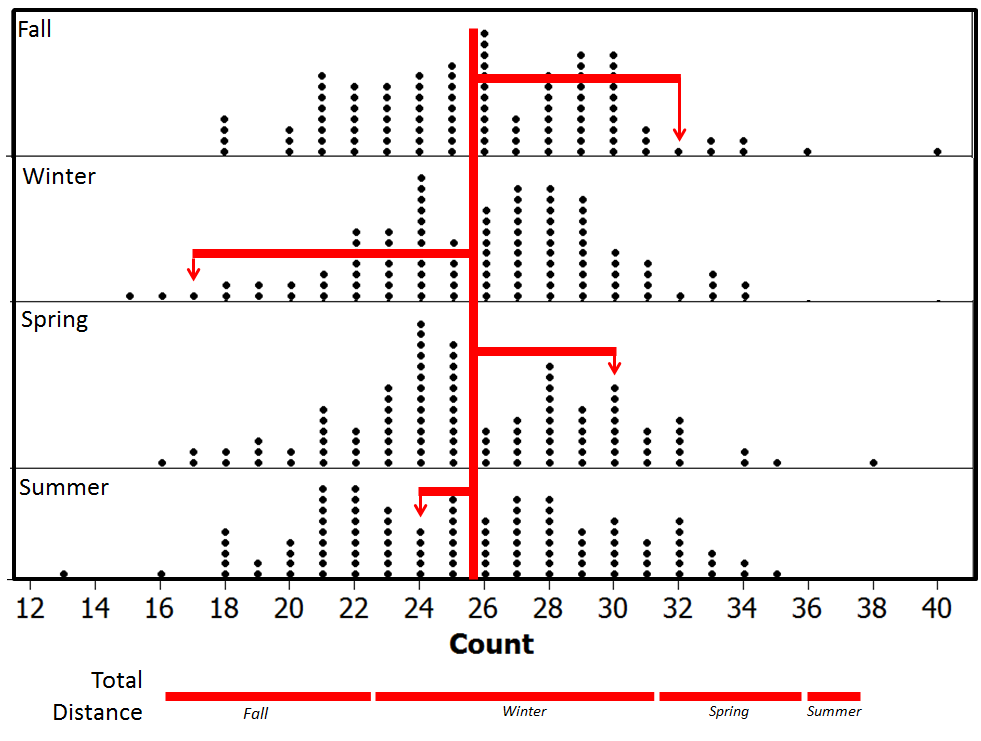
The math for determining familywise and maximum error rates when multiple p-values are used to make a decision.

|  |  |
| --- | --- |
| * Familywise Error Rate     where *k* = # of tests being considered   * Maximum Error Rate (Boole’s Inequality) | Table Showing Possible Error Rates |

*Source*: Wiki page on Multiple Comparisons; <http://en.wikipedia.org/wiki/Multiple_comparisons>

Measuring Distance between Observed and Expected with Several Categories

As mentioned above, having multiple p-values is problematic when a single decision is to be made regarding a single research question. To overcome this problem, the distance from the Observed to the Expected Value is what is considerd in our formal statistical test. This is shown below.



Compute the distance from the Observed to the Expected for the Spring and Summer seasons.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| U of MN | Season | | | | Total |
| Fall | Winter | Spring | Summer |
| Observed | 32 | 17 | 30 | 24 | 103 |
| Expected | 25.75 | 25.75 | 25.75 | 25.75 | 103 |
| Distance | 32 – 25.75 = 6.25 | 17 – 25.75 = -8.75 |  |  |  |

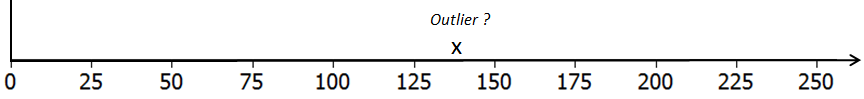
Questions

1. Add the Distance row in the table above. What is the total distance? Does this value make sense for total distance? How might we overcome this issue?

Taking the square of each distance is shown in the table below. This is done so that the negative distances do not cancel out the positive distances. The absolute values could have been used as well to get rid of the negatives; however, squaring each distance is used here.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| U of MN | Season | | | | Total |
| Fall | Winter | Spring | Summer |
| Observed | 32 | 17 | 30 | 24 | 103 |
| Expected | 25.75 | 25.75 | 25.75 | 25.75 | 103 |
| Distance | 6.25 | -8.75 | 4.25 | -1.75 | 0.00 |
| Distance2 | 39.06 | 76.56 | 18.06 | 3.06 | **136.74**  **≈ 137** |

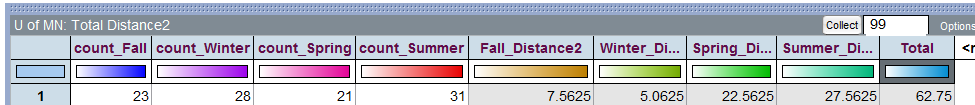
The total squared distances summed up across all four seasons is about 137. We cannot determine whether or not 137 is an outlier using our previous graphs. The previous graphs considered each season individually. Our new measure is the squared distance between the Observed and Expected summed over four seasons. A new graph,a single graph, will need to be created in Tinkerplots to determine whether or not 137 is an outlier.



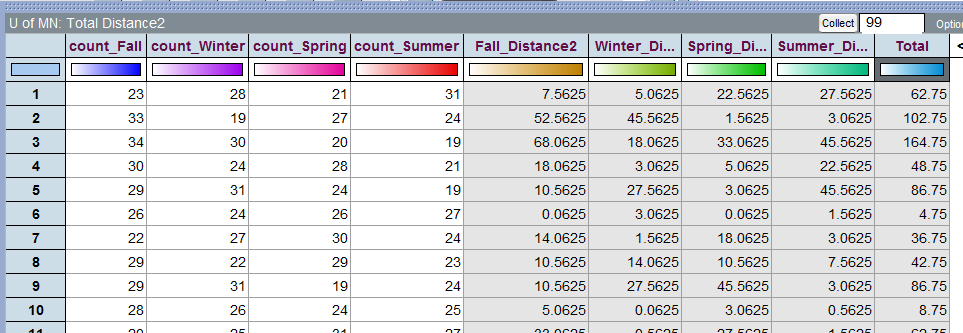
Questions

1. What would a value of 0 imply on the above number line? Explain why a value less than 0 is not possible when the distances are squared and summed across the categories.
2. What would a large value imply? Is this evidence for or against the original research question? Explain.
3. When squared distances are computed and summed across all categories, the appropriate test is one-tailed right. Explain why this is the case.

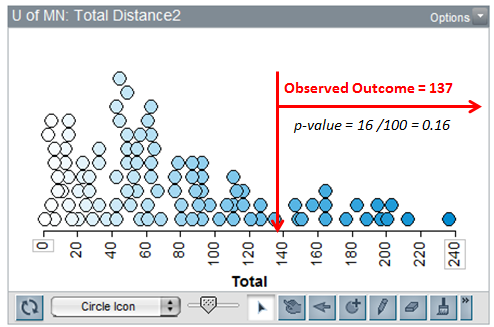
In Tinkerplots, a formula can be used in the History table to compute the squared distance between the simulated outcome for a single trial and the expected for each season. These squared distance values are then summed across the four seasons. The total squared distance for the 1st trial is 62.75, this is shown here.



When additional trials are done in Tinkerplots, these distances and total are computed automatically for each trial. The outcomes from the first 10 trials are shown here.



A graph of the total squared distances from 100 trials done in Tinkerplots. The p-value is determined using the proportion of dots greater than or equal to 137, the “observed outcome” from the study.

.

Questions

1. What is an approximate p-value from the above graph? What is the appropriate statistical decision for our research question?

3.3: Testing Proportions that are Not Equal Across Categories

Example 3.3.1: The Minnesota Student Survey (MSS) is a survey administered every three years to 6th-, 9th- and 12th-grade students and also is offered to students in area learning centers and to youth in juvenile correctional facilities. The survey is an important vehicle for youth voice. School district leaders and educators, local public health agencies and state, community and social services agencies use the survey results in planning and evaluation for school and community initiatives and prevention programming.

Questions are asked related to both the home and school life of students; topics include family relationships, feelings about school, substance use, wellness activities, and more. Participation in the survey is voluntary, confidential and anonymous.

For the analysis here, we will consider Question # 105 from this survey. Data has been collected for Fillmore County which is in Southeastern Minnesota. The population of Fillmore County is 20,866 and consists of several small rural communities.

|  |  |
| --- | --- |
| Question #105 from MN Students Survey | Fillmore County is in  Southeastern Minnesota |

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



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

The survey is completed by students in Grade 6, 9, and 12. Information regarding the historical patterns for Grade 6 students from across the state of Minnesota is given here and will be used for comparisons.

Historical Patterns for Grade 6 Students Across the State of Minnesota

* + About 3 out of 4 students in Grade 6 respond to the third part of this question (i.e. “smoking marijuana once or twice a week” ) with “Great Risk”
  + A very small percentage, only about 1%, respond to the third part of this question with “No Risk”
  + The remaining students typically divide themselves between “Slight Risk” and “Moderate Risk” when responding to the third part of this question.

|  |  |
| --- | --- |
| **Fillmore County Marijuana Case Study** | |
| Research Question | Is there evidence to suggest that Grade 6 students from Fillmore County deviate from historical patterns of the marijuana portion of Question 105? |
| Testable Hypothesis | Ho: Fillmore County Grade 6 students do not deviate from historical patterns  HA: Fillmore County Grade 6 students deviate from historical patterns. |
| Parameters | The four parameters of interest are defined as follows:  = the probability of a Grade 6 student from Fillmore County will respond   to the marijuana portion of this question with No Risk  = the probability of a Grade 6 student from Fillmore County will  respond to the marijuana portion of this question with Slight Risk  = the probability of a Grade 6 student from Fillmore County will  Respond to the marijuana portion of this question with  Moderate Risk  = the probability of a Grade 6 student from Fillmore County will  respond to the marijuana portion of this question with Great Risk |
| Rewrite of Hypotheses | Identify the appropriate proportions for each category |

Consider the following table. The first row of this table contains the Observed Outcomes for Grade 6 students from Fillmore County (Male and Female tallies were combined) and the second row contains the Expected Outcome (i.e. Anticipated Outcome under the null hypothesis) for each of the possible survey responses.

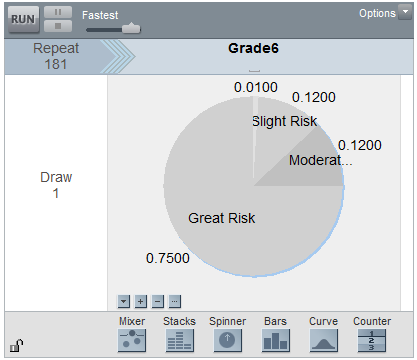
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Type of Outcome | Survey Responses for Fillmore County, 6th Grade | | | | Total |
| No Risk | Slight Risk | Moderate Risk | Great Risk |
| *Observed* | 9 | 9 | 20 | 143 | 181 |
| *Expected* | 1.81 | 21.72 | 21.72 | 135.75 | 181 |

Questions

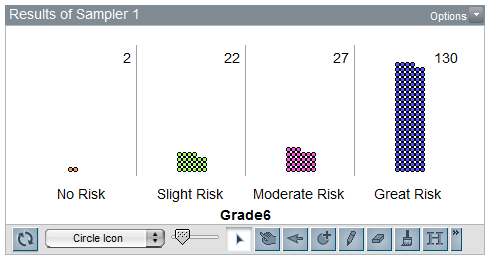
1. The value in the first row and second column is 9 (i.e. *Observed* / Slight Risk). Explain where this number came from. What does this value represent?
2. What does the Total value for the Observed row represent?
3. The value for the second row and second column is 21.72 (i.e. *Expected* / Slight Risk). Explain where this number came from. How was it computed? What does this value represent?
4. Why is the Expected Value for Great Risk so much higher than the others? In the context of this problem, why would a county health nurse be excited to have such a high number in this category? Explain.
5. Your friend computes the following percentages: No Risk: 9/181 ≈ 5%; Slight Risk: 9/181 ≈ 5%; Moderate Risk: 20/181 ≈ 11%; and Great Risk: 143/181≈ 79%. Your friend then makes the following statement, “There is enough evidence for the research question because these percentages are different than the historical percentages (i.e. No Risk = 1%, Slight Risk = 12%, Moderate Risk = 12%, and Great Risk = 75%).” Why is this a statistically incorrect statement? Explain.

Once again, in an effort to understand the amount acceptable deviation from the expected, a simulation can be run in Tinkerplots.

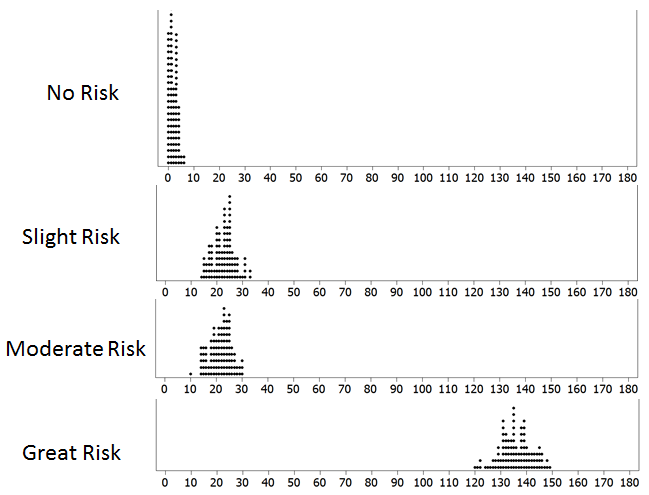
Spinner setup in Tinkerplots under the assumption that Fillmore County Grade 6 students follow historical patterns.



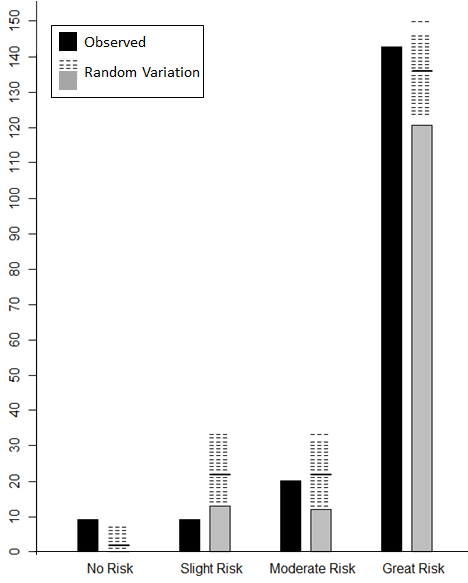
Outcome from a single trial. Notice the large proportion of outcomes for the Great Risk category.



A graph showing the outcomes from 100 trials.



The following graph is a slight variation of the above. This bar graph shows the acceptable amount of random variation under the null hypothesis for each category.



Determine whether or not the observed outcome for each category would be considered an outlier.

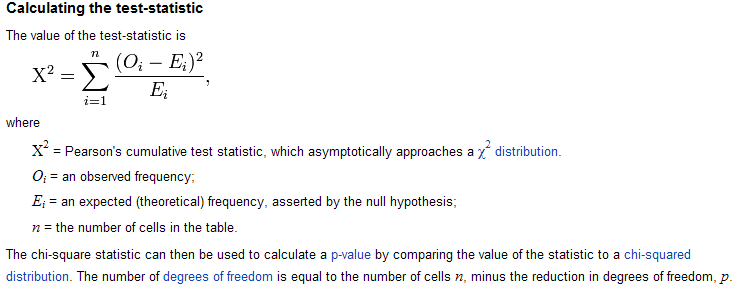
|  |  |  |  |
| --- | --- | --- | --- |
| Survey Outcome | Outlier? | | |
| Yes | No | Maybe |
| No Risk | X |  |  |
| Slight Risk |  |  |  |
| Moderate Risk |  |  |  |
| Great Risk |  | X |  |

Consider once again the distance between the Observed and the Expected outcome for each of the possible choices for this question.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Survey Responses for Fillmore County, 6th Grade | | | | Total |
| No Risk | Slight Risk | Moderate Risk | Great Risk |
| *Observed (i.e.O)* | 9 | 9 | 20 | 143 | 181 |
| *Expected (i.e. E)* | 1.81 | 21.72 | 21.72 | 135.75 | 181 |
| Difference = (O - E) | 7.19 | -12.72 | -1.72 | 7.25 | 0 |

We discussed earlier the fact that we need to square the differences (i.e. distances) before summing because the positive and negative values cancel each other out. Upon careful inspection of these differences, there is another problem that is more substantial that needs to be addressed. In particular, notice that the Great Risk difference is 7.25 and the No Risk difference is slightly smaller at 7.19. However, when we identified the outliers above, we said 9 was an outlier for No Risk, but 143 was not an outlier for Great Risk. The discrepancy is a problem and can be summarized as follows -- when determining extremeness, we need to measure the distance or squared distance AND take the scale into consideration. Thus, for each category, the following is computed. These quantities are summed across all categories and the resulting value is called the **Test Statistic**.

Wiki Entry for Test Statistic for this test

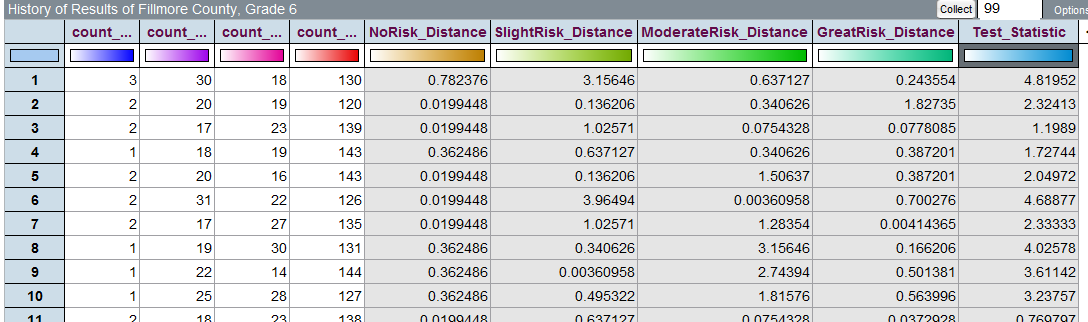


|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Survey Responses for Fillmore County, 6th Grade | | | | Total |
| No Risk | Slight Risk | Moderate Risk | Great Risk |
| *Observed (i.e.O)* | 9 | 9 | 20 | 143 | 181 |
| *Expected (i.e. E)* | 1.81 | 21.72 | 21.72 | 135.75 |  |
| Difference = (O - E) | 7.19 | -12.72 | -1.72 | 7.25 |
|  | 51.7 | 161.8 | 2.96 | 52.56 |
|  | 28.56 | 7.45 | 0.14 | 0.39 | **36.54** |

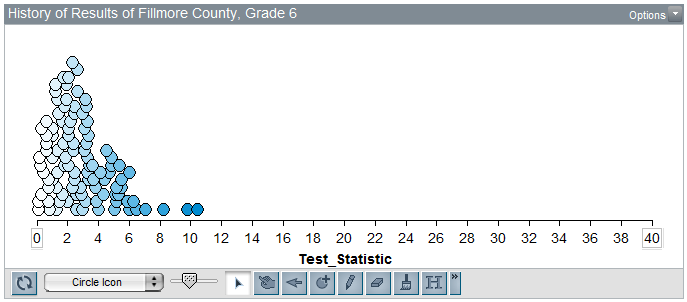
The test statistics for this analysis would be

Test Statistic: \_\_36.54\_\_

Akin to what was done before, now we must determine whether or not the Test Statistic from the observed data (i.e. 36.54) would be considered an outlier. Tinkerplots can be used to compute the Test Statistic over repeated samples under the null hypothesis. The outcomes from the first 10 trials is shown here (see far right column in table below). It appears 36.54 would be considered an outlier.



A graph of the 100 trials.



Questions

1. The Test Statistic for the observed data is 36.54. Would this value be considered an outlier in this reference distribution? Explain.
2. Consider the formula for the test statistic. What would a value near 0 imply? What would a large value imply? Explain.

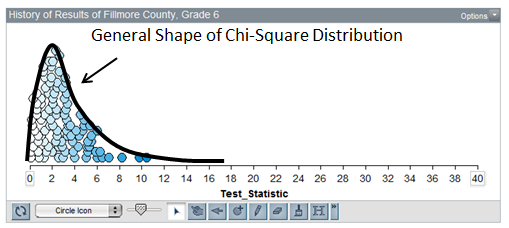
Compute the p-value, make a decision, and write a final conclusion for the original research question.

|  |  |
| --- | --- |
| **Fillmore County Marijuana Case Study** | |
| Research Question | Is there evidence to suggest that Grade 6 students from Fillmore County deviate from historical patterns of the marijuana portion of Question 105? |
| Testable Hypothesis | Ho: Fillmore County Grade 6 students do not deviate from historical patterns  HA: Fillmore County Grade 6 students deviate from historical patterns. |
| P-Value | p-value = the proportion of test statistics that would be as extreme or more extreme than the observed test statistic 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 for the Director of the Minnestoa Department of Education that address the original research question. |

Comments The Wiki entry above mentions the Chi-Square Distribution. This is the name of the reference distribution for the test being performed here. This is different than the Binomial distribution presented in Chapter 2. Consider some of the differences discussed here.

* Unlike the binomial distribution, the chi-square distribution is not based on counts and is often skewed to the right.



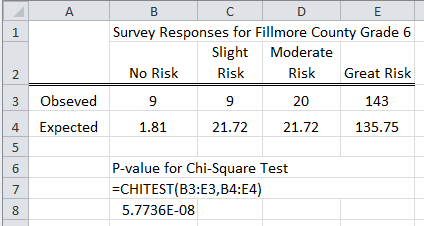
* The number of categories is taken into consideration through a quantity called the degree-of-freedom, typically referred to as the **df value**.

*df* = # Categories – 1

* Excel can be used to determine appropriate critical values and p-values. Each of these quantities depend on the degrees-of-freedom (i.e. df) value.

|  |  |
| --- | --- |
| Critical Values | Computing the a p-value |

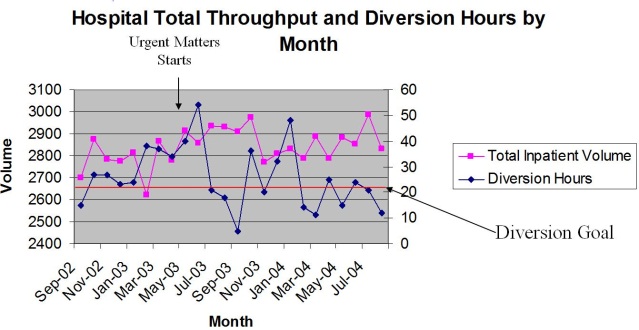
* The CHITEST() function in Excel can be used to obtain the p-value directly. This function takes two arguments – the observed outcomes and the expected outcomes. Thus, before using this function, you must compute the appropriate expected values for each category.



3.4: The Chi-Square Goodness-of-Fit Test in Excel

The name of the test completed in the previous section is called a **Goodness-of-Fit -- Chi Square Test**. This name is often shortened to Chi-Square test; however, there are a variety of test that use the Chi-Square distribution as it’s reference distribution; thus, statisticians refer to this particular type of Chi-Square Test as a Goodness-of-Fit test.

Example 3.4.1: Consider the following study done at Boston Medical Center. This study was centered around the scheduling of elective surgeries over a two year period. Scheduling surgeries is important to quality patient care and is an important health care management issue as well. In particular, consistency in the scheduling of the surgeries results in less cancellation and an appropriate level of staffing is more obtainable. This medical center made a significant change in their approach to scheduling surgeries between year 1 and year 2.



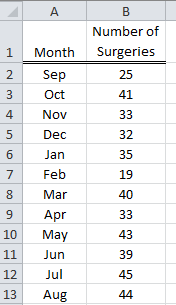
Consider the following data extracted from the above graph.

|  |  |
| --- | --- |
| Total inpatient volume in Year 1 | Total inpatient volume in Year 2 |

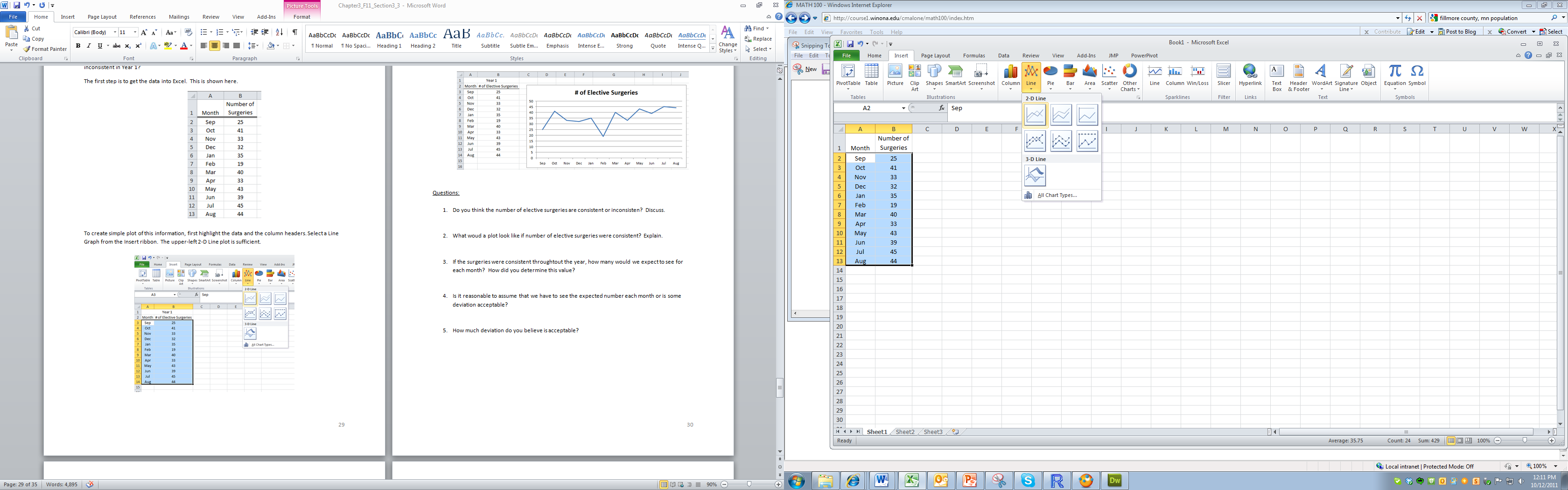
Consider the following research question. Define at least one of the parameter of interest in this case study.

|  |  |
| --- | --- |
| **Boston Medical Center Case Study** | |
| Research Question | Is there enough statistical evidence to suggest that total inpatient volume was inconsistent in Year 1? |
| Testable Hypothesis | Ho: Total inpatient volume at Boston Medical Center is consistent  HA: Total inpatient volume at Boston Medical Center is not consistent |
| Parameters | There are 12 parameters that need to be considered for this investigation --  , , , …,  Give a complete definition of one of the parameters, say .  = |
| Rewrite of Hypotheses | The null and alternative hypotheses are stated here. |

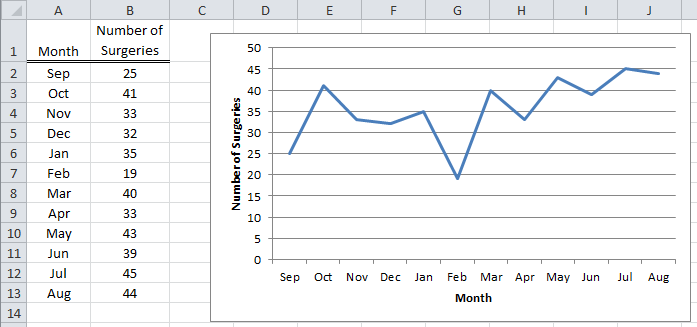
The observed outcomes in Excel.



To create simple line chart of this information, first highlight the data and the column headers. Select a Line Graph from the Insert ribbon. The upper-left 2-D Line plot is sufficient.



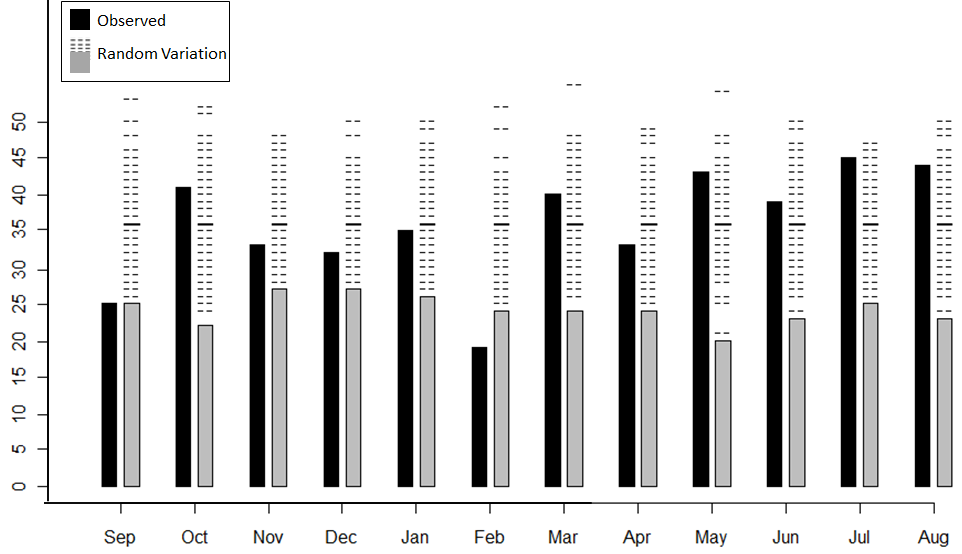
The following line plot is created.



Questions:

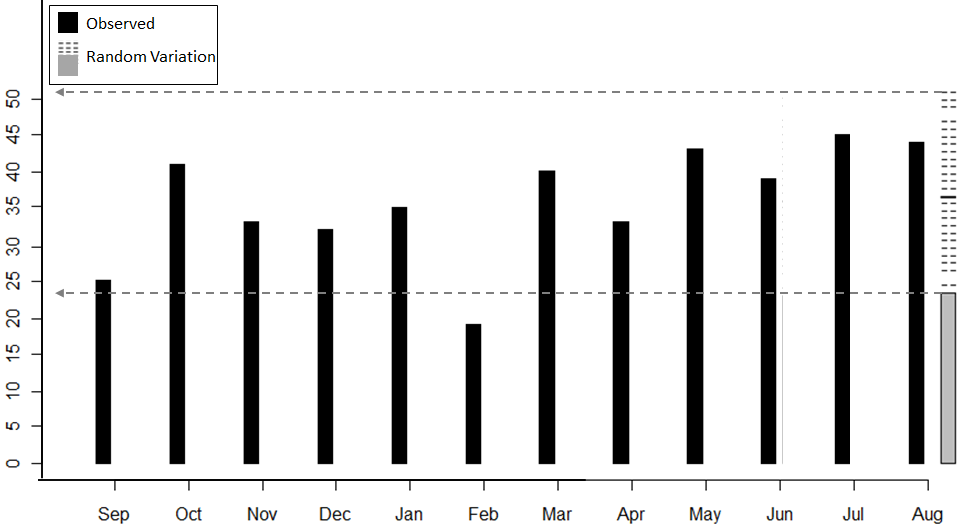
1. How many surgeries occurred in Year 1? That is, what is the total number of surgeries in this investigation? Discuss.
2. What is the expected number of surgeries, for each month, if the surgeries are occuring equally throughout the year? Explain how you obtained this value. Sketch the expected value for each month on the above line graph.
3. Do you believe the variation seen in the observed data is within the boundaries of random variation (i.e. within the ‘margin-of-error’)? Explain.

Consider the following plot that shows the amount of random variation under the assumption that surgeries are occuring equally.

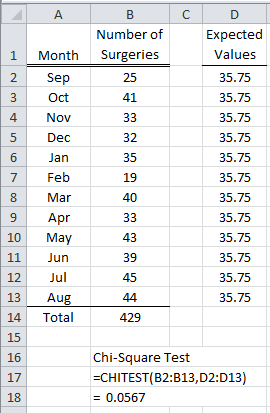


Questions:

1. For each month, the expected number (i.e. the bold line amonst the dashed lines) is the same. Why is this the case for this case study?
2. The amount of acceptable random variation is about the same for each month. Explain why this should be the case.
3. Using the above plot, which months would be considered outliers? How did you make this determination?
4. Your friend makes the following statement, “It’s rediculous to do repeated samples for each month . You can use the outcomes from one month, say Aug, for all the others to determine whether or not a month is an outlier. A statistician would agree with this. Why?

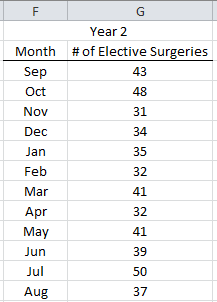


Getting our test done in Excel…



|  |  |
| --- | --- |
| **Boston Medical Center Case Study** | |
| Research Question | Is there enough statistical evidence to suggest that total inpatient volume was inconsistent in Year 1? |
| Testable Hypothesis | Ho: Total inpatient volume at Boston Medical Center is consistent  HA: Total inpatient volume at Boston Medical Center is not consistent |
| P-Value | 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 for the nurse in charge of staffing the inpatient surgery rooms at Boston Medical Center. |

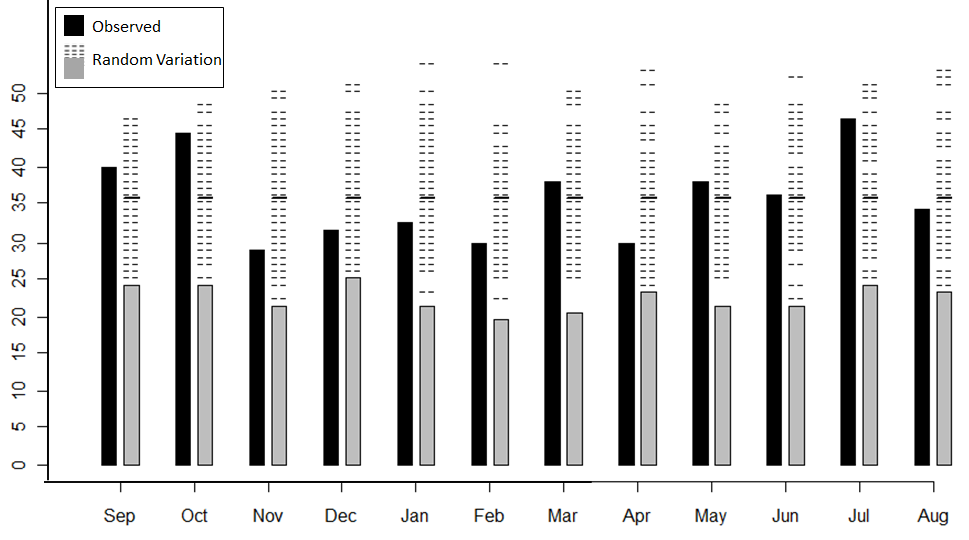
Example 3.4.2: Reconsider the data from the Boston Medical Center. Rerun the analysis done on Year 1 using data from Year 2.



|  |  |
| --- | --- |
| **Boston Medical Center Case Study** | |
| Research Question | Is there enough statistical evidence to suggest that total inpatient volume was inconsistent in Year 2? |
| Testable Hypothesis | Ho: Total inpatient volume at Boston Medical Center is consistent  HA: Total inpatient volume at Boston Medical Center is not consistent |
| Rewrite of Hypotheses | The null and alternative hypothesis with the parameters. |
| P-Value | Use Excel to obtain the appropriate p-value.    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 for the nurse in charge of staffing the inpatient surgery rooms at Boston Medical Center. |

Questions:

1. Consider the following graphic that show the amount of acceptable random varition for each month for the Year 2 investigation. Does this graph support or refute the conclusion reached above? Explain.



1. As mentioned in Example 3.4.1, the Boston Medical Center made a change in their scheduling of elective surgeries between Year 1 and Year 2. Did this change have the desired effect? That is, was total patient volume more consistent in Year 2 than in Year 1? Compare and contrast the p-value from Year 1 to that of Year 2 to support your answer.

Example 3.4.3: For this example, consider the following continuation of the data from the Boston Medical Center. For insurance companies it is profitable to delay elective surgeries until January because deductibles have not yet been met. For this example, we will attempt to investigate whether or not insurance companies take advantage of this fact? If this is happening, then we’d expect January to have a higher than expected number of elective surgeries. In the real world, this issue is complicated by the fact that a medical center typically tries to reduce the number of surgeries around the Holidays especially around Christmas and New Year’s.

Suppose a typical level of reduction in elective surgeries at Boston Medical Center is 20% for December. As a result, we would expect January to have a 20% higher rate in the number of elective surgeries. Use the data from Year 1 to determine whether or not insurance companies appear to be adversely affecting when surgeries take place by forcing such surgeries to take place in January instead of December.

|  |  |
| --- | --- |
| **Insurance Deductibles / Boston Medical Center Case Study** | |
| Research Question | Is there statistical evidence to suggest that insurance companies are adversely affecting when elective surgeries are taking place (i.e. total patient volume) at Boston Medical Center? |
| Testable Hypothesis | Ho: Insurance companies have no impact on when surgeries are taking place;   that is, the percentages follow the trends stated by the Boston Medical   Center  HA: Insurance companies have an impact on when surgeries are taking place;   that is, the percentages are different than those stated by the Boston   Medical Center |
| Rewrite of Hypotheses | The null and alternative hypotheses are given here. Notice the proportions for December and January have been appropriately adjusted according to the information provided by the Boston Medical Center. |

|  |  |
| --- | --- |
| P-Value | Enter the Number of Surgeries into Excel for Year 1. Compute the appropriate expected values. Use the CHITEST() function to obtain the p-value for this 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 for a lawyer whose client is being asked to cover charges that are typically covered under their deductable for a recent elective surgery at Boston Medical Center. |

Question:

1. Consider the following graphic that show the amount of acceptable random varition for each Consider the conclusion given above. Before one claims that insurance companies are unethical by making patients wait until January for their surgeries, we should consider the last column, provided on far right of the above screen shot from Excel. This column contains the cell contribution towards the test statsitic. The larger this value is, the more that cell (or categories) is contributing to a small p-value.



February, September, and July are contibuing the most to having a low p-value (i.e. or high test statistic) because their cell contribution value is highest. Does this fact, support or refute the notation that insurance compaines are being unethical and making pateients wait for elective surgeries until January? Explain.

Example 3.4.4: Consider the following data from a survey recently done by Nationwide.com, an auto insurance company. Their website mentions that 1,008 adults were selected to participate in this survey. Consider the following question from this survey.

|  |
| --- |
| Survey Question: Would a law making it illegal to text or talk on a cell phone while driving cause you to change your own behavior, personally? |

Their definitions for generations:

* Gen Y: 21 – 32 years old
* Gen X: 33 – 44 years old
* Boomers: 45-63 years old, and
* Seniors: 64+ years old

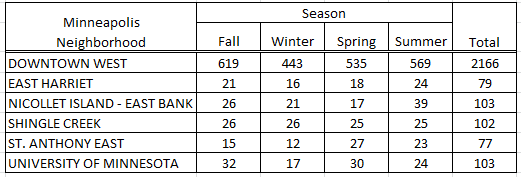
Consider the following typical percentages for each type of response from this survey question.

|  |  |
| --- | --- |
| Response | Reference Values |
| Yes | 41% |
| No, will con’t to use cell phone | 9% |
| No, already doesn’t use cell phone | 49% |
| Don’t Know / Refused | 1% |

|  |  |
| --- | --- |
| **Nationwide.com Case Study** | |
| Research Question | Are the responses from Gen Y statistically different than the specified reference values? |
| Testable Hypothesis | Ho: Proportions for Gen Y are the same as the reference values  HA: Proportions for Gen Y are the different than the reference values |
| Parameters | The four parameter of interest  = the probability that a Gen Y survey respondent will answer Yes  = the probability that a Gen Y survey respondent will answer No,  will continue to use cell phone  = the probability that a Gen Y survey respondent will answer No,  already doesn’t use cell phone  = the probability that a Gen Y survey respondent will answer Yes |
| Rewrite of Hypotheses | The null and alternative hypothesis with the parameters. |
| P-Value | Frist, obtain the appropriate expected values for each outcome. Use Excel to obtain the p-value for the goodness-of-fit 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 that would be appropriate for an indvidual working at Nationwide Auto Insurance that is supposed to write a press release regarding the analysis of this question to the public. |

Example 3.4.5: Consider the Minneapolis Police Crime dataset which is available on our course website. The table below contains the counts for total crimes across the four seasons for six different neighborhoods in Minneapolis.

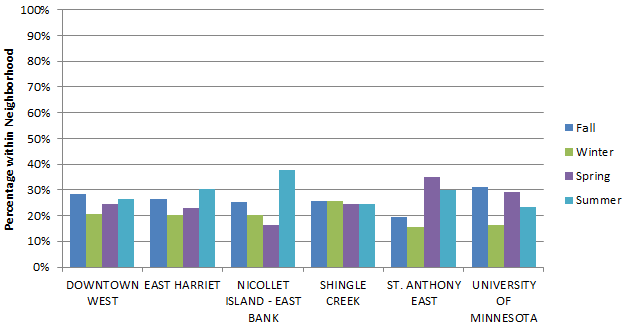


For this example, you are to pick one of these neighborhoods or any other neighborhood from this dataset and complete a goodness-of-fit test to statistically determine whether or not crimes are occurring equally throughout the year for that neighborhood.

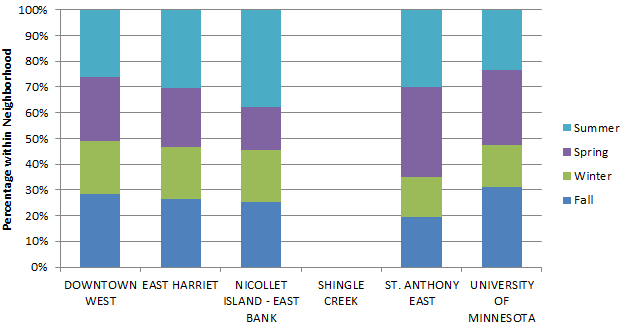
|  |  |
| --- | --- |
| **Which neighborhood did you choose? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Case Study** | |
| Research Question | Write the appropriate research question for your neighborhood. |
| Testable Hypothesis | Write the appropriate null and alternative hypothesis for your investigation.  Ho:  HA: |
| Parameters | How many parameters are relevant for your investigation? Carefully define at least one of the parameters. |
| P-Value | Enter the appropriate observed data in Excel, compute the expected values, and complete the Chi-Square Goodness-of-Fit test in Excel using the CHITEST() function.  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 for the Chief of Police for the Minnesapolis Police Department regarding your investigation. |

Questions:

1. Consider the following graph. Consider the neighborhood you decided to investigate above. Does this graph support or refute the conclusion that you reached by completing the above test? Explain.



1. Consider the graph above. A variation of this side-by-side bar chart is give below. In this plot, the bars for each neighborhood are stacked on top of one another. This type of graph is called a 100% Stacked Column Chart in Excel (statisticians call this type of plot a **mosaic plot**).
   1. Plot the missing information for Shingle Creek in this plot.



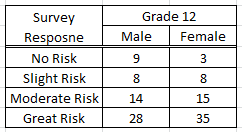
* 1. Pick any neighborhood. What would be appearance of the stacked bar if the crimes were occurring equally across the four seasons? What would be the appearance if the crimes were not occurring equally? Explain.
  2. Consider the lowest box for Downtown West (i.e. the blue box). The height of this box is about 28% or 29%. How did Excel compute this number? In particular, what values did it use to determine this percentage? Explain.
  3. Which of these neighborhoods appear to have similar crime patterns across the four seasons of the year? Discuss.

Example 3.4.6: For this example, consider the data for Grade 12 students from Fillmore County who have completed the Minnesota Student Survey.



|  |  |
| --- | --- |
| Historical Trends for Grade 12 Students Across Minnesota | |
| Males | Females |
| Historical Patterns for Grade 12 Males   * About 30% respond with Great Risk * The remaining respondents typically split evenly across the remaining categories (i.e. No Risk / Slight Risk / Moderate Risk) | Historical Patterns for Grade 12 Females   * About 50% respond with Great Risk * Only about 10% of Grade 12 Females believe there is No Risk * The remaining 40% is split evenly across Slight Risk and Moderate Risk |

Use the following data to complete a Goodness-of-Fit test for Grade 12 Males from Fillmore County. Compare this data against the historical patterns quoted above.



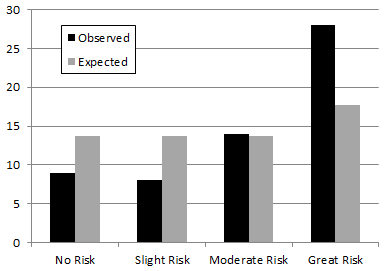
|  |  |
| --- | --- |
| **Grade 12 Males from Fillmore County Case Study** | |
| Research Question | Write the appropriate research question for your neighborhood. |
| Testable Hypothesis | Write the appropriate null and alternative hypothesis for your investigation.  Ho:  HA: |
| Parameters | How many parameters are relevant for your investigation? Carefully define at least one of the parameters. |
| P-Value | Enter the appropriate observed data in Excel, compute the expected values, and complete the Chi-Square Goodness-of-Fit test in Excel using the CHITEST() function.  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 a high school principle from Fillmore County. |

Repeat the analysis completed above for Grade 12 Females from Fillmore County.

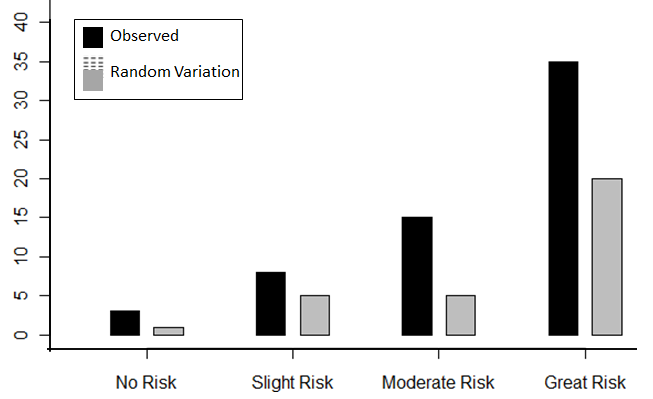
|  |  |
| --- | --- |
| **Grade 12 Females from Fillmore County Case Study** | |
| Research Question | Write the appropriate research question for your neighborhood. |
| Testable Hypothesis | Write the appropriate null and alternative hypothesis for your investigation.  Ho:  HA: |
| Parameters | How many parameters are relevant for your investigation? Carefully define at least one of the parameters. |
| P-Value | Enter the appropriate observed data in Excel, compute the expected values, and complete the Chi-Square Goodness-of-Fit test in Excel using the CHITEST() function.  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 a high school principle from Fillmore County. |

Questions:

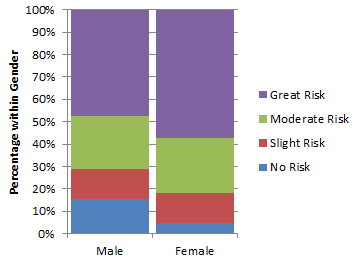
1. Above, we learned that the pattern for Grade 12 Males from Fillmore County was statistically different than the historical patterns (p-value = 0.018). Consider the following graph that shows the observed counts and expected counts for Grade 12 Males from Fillmore County. From this graph, are males doing “better” or “worse” than the historical trends? Explain.



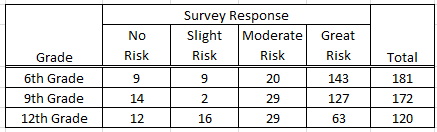
1. Above, we learned that the pattern for Grade 12 Females from Fillmore County was not statistically than the historical patterns (p-value = 0.228). Consider the following plot which is missing the expected counts and the dashed lines representing the random variation.
   1. Add a bold line representing the expected count for each outcome.
   2. Next, consider the fact that the p-value was bigger than 0.05. Sketch an appropriate amount of random variation for each outcome. Explain why the observed bars would \*not\* be considered outliers against this random variation.



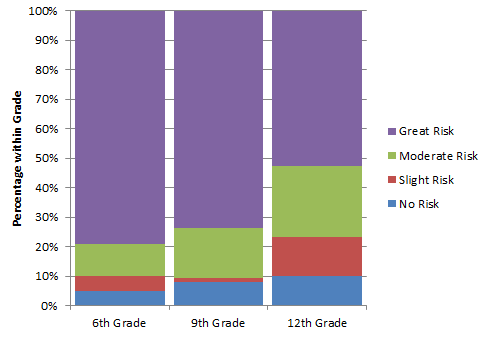
1. Consider the following stacked bar chart that could be used to compare Grade 12 Males to Grade 12 Females from Fillmore County. Discuss any noticeable differences in patterns between the males and females. Are there any similarities? Discuss.



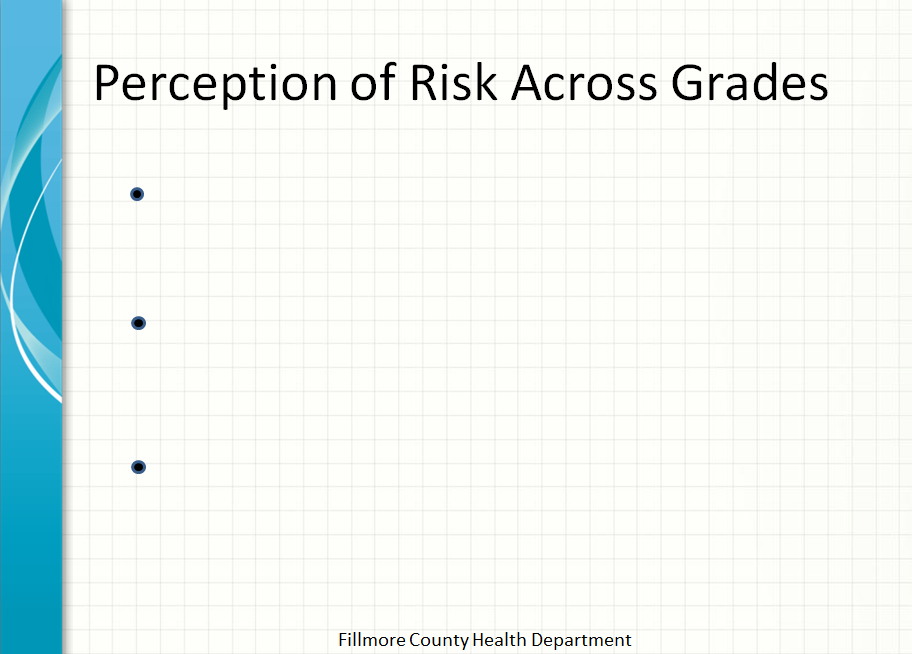
1. Consider the number of students from Fillmore County that responded to each survey response across the three grade levels (i.e. Grade 6, 9, and 12).



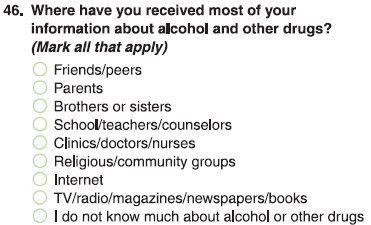
The following stacked bar chart show the trends present across the three grade levels.



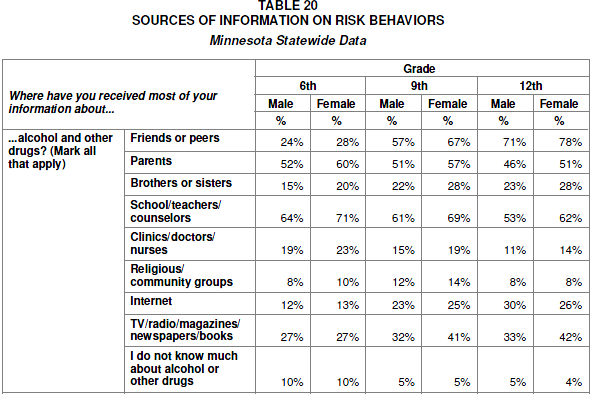
Write three statements to assist the Director of the Fillmore County Health Department in preparing a PowerPoint presentation that summarizes what is learned by looking at this data and/or graph.



Comment: Some survey questions are more difficult to analyze than others. On such example is “Mark All that Apply” questions. Question #46 from the most recent Minnesota Student Survey is given here as an example.



Suppose the goal is to compare data for Grade 12 Males from Winona County to the Minnesota Statewide Data. The trends for the statewide data are given here.



Questions:

1. The percentages for Grade 12 Males do not add up to 100%? Explain why this might be the case for a “Mark All that Apply” type question.
2. Why is it \*not\* possible to setup a spinner in Tinkerplots for an investigation of Grade 12 Males for this question? Explain.