INFERENCE FOR A CATEGORICAL VARIABLE WITH MORE THAN TWO CATEGORIES

The analyses we have completed up to this point in the semester were for a single categorical variable with only <u>two</u> outcomes. For example, in the helper/hinderer study, the babies were choosing either one toy or the other. In the gender discrimination example, each employee selected for management was either male or female. Next, we'll consider problems involving a single categorical variable which has more than two categories.

Example 2.7: The Minneapolis Police Department posts regular updates on crime statistics on their website. A colleague of mine has collected this data for the past few years on all neighborhoods in Minneapolis. A portion of the data set and the precinct map are shown below.

<table>
<tr><td colspan="15" align="center">Minneapolis Crime Statistics – a snippet of the data</td><td colspan="2" align="center">Precinct Map</td></tr>
</table>

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Fiscal Year	Season	Month	Month ID	Precinct	Neighborhood	Total	Homicide	Rape	Robbery	AggAssault	Burglary	Larceny	AutoTheft	Arson
2	Current	Summer	Aug	1	5	ARMATAGE	5	0	0	0	0	2	3	0	0
3	Current	Summer	Aug	1	2	AUDUBON PARK	13	0	0	2	0	4	7	0	0
4	Current	Summer	Aug	1	3	BANCROFT	17	0	0	1	0	8	6	2	0
5	Current	Summer	Aug	1	2	BELTRAMI	4	0	0	0	0	1	3	0	0
6	Current	Summer	Aug	1	2	BOTTINEAU	10	0	0	2	0	5	2	0	1
7	Current	Summer	Aug	1	3	BRYANT	9	1	0	0	2	0	4	2	0
8	Current	Summer	Aug	1	4	BRYN - MAWR	16	0	0	0	0	3	11	2	0
9	Current	Summer	Aug	1	4	CAMDEN INDUSTR	2	0	0	0	0	0	2	0	0
10	Current	Summer	Aug	1	5	CARAG	36	0	0	0	3	13	19	1	0
11	Current	Summer	Aug	1	5	CEDAR - ISLES - DE	13	0	0	0	0	2	9	2	0
12	Current	Summer	Aug	1	1	CEDAR RIVERSIDE	34	0	0	1	4	2	24	3	0
13	Current	Summer	Aug	1	3	CENTRAL	45	0	2	6	4	22	10	1	0
14	Current	Summer	Aug	1	4	CLEVELAND	23	0	1	1	4	13	2	2	0
15	Current	Summer	Aug	1	2	COLUMBIA PARK	6	0	0	0	0	1	4	1	0
16	Current	Summer	Aug	1	2	COMO	19	0	1	1	0	5	8	4	0
17	Current	Summer	Aug	1	3	COOPER	17	0	0	0	0	5	11	1	0
18	Current	Summer	Aug	1	3	CORCORAN	28	0	1	3	3	6	13	2	0
19	Current	Summer	Aug	1	3	DIAMOND LAKE	10	0	0	0	1	3	3	3	0
20	Current	Summer	Aug	1	1	DOWNTOWN EAST	13	0	1	1	0	2	9	0	0
21	Current	Summer	Aug	1	1	DOWNTOWN WES	192	0	0	15	8	3	158	8	0
22	Current	Summer	Aug	1	5	EAST HARRIET	8	0	0	0	0	3	5	0	0
23	Current	Summer	Aug	1	5	EAST ISLES	29	0	0	0	0	4	23	2	0
24	Current	Summer	Aug	1	3	EAST PHILLIPS	54	0	2	5	6	6	31	4	0
25	Current	Summer	Aug	1	5	ECCO	19	0	1	0	0	5	12	1	0

Source: http://www.minneapolismn.gov/police/crime-statistics/
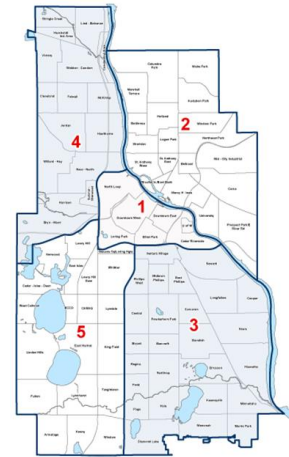
Suppose the police chief for Precinct #2 has received a complaint from a permanent resident who lives in a neighborhood near the University of Minnesota. This resident has asked for additional patrol to take place in his neighborhood as he believes that crime rates vary over the course of the year.

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 we will 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 Observed Data

Crimes classified as Murder, Rape, Robbery, Aggravated Assault, Burglary, Larceny, Auto Theft, and Arson are used in reporting the **total number of crimes**. The Minneapolis Police Department reported that a total of 103 crimes occurred in the University of Minnesota neighborhood in this particular year, and the table below shows the breakdown of these counts (and percentages) across season.

	Season				Total
	Fall	Winter	Spring	Summer	
Observed Count	32	17	30	24	103
Observed Percentage	31.1%	16.5%	29.1%	23.3%	100%

Note that the police chief may consider conducting a hypothesis test to investigate the concern expressed by his resident. We can formalize hypotheses as shown below.

Minneapolis Crime Case Study	
Research Question	Do these data provide enough evidence for the police chief to conclude that crime patterns in the University of Minnesota neighborhood differ over the four seasons of the year?
Parameters of Interest	The four parameters of interest are defined as follows: π_{fall} = the probability of a crime occurring in the fall π_{winter} = the probability of a crime occurring in the winter π_{spring} = the probability of a crime occurring in the spring π_{summer} = the probability of a crime occurring in the summer
Hypotheses	H_o: Crimes are equally dispersed over the four seasons H_a: Crimes are not occurring with equal probability over the four seasons

A Simulation Study

The approach we will take to carry out this hypothesis test is very similar to what we have done previously.

- We will *assume* the crime patterns are occurring equally across the four seasons (i.e., that the null hypothesis is true) and then get a good idea of what outcomes we would expect to see if this were really the case.

- Then, we will check to see if the observed outcomes given in the data are consistent (or inconsistent) with what we expected to see under the null hypothesis.

- If the observed data are inconsistent with the outcomes expected under the null, then we will have sufficient statistical evidence to say crime rates vary across the four seasons.

Questions:

1. Find the expected number of total crimes for each season under the assumption that crimes are occurring equally over the four seasons. How did you obtain these values?

	Season				Total
	Fall	Winter	Spring	Summer	
Expected Count					103

2. Recall the *actual* crime statistics for the University of Minnesota neighborhood for that particular year. Even if in the long run crime patterns don't really differ across season, do we anticipate these observed counts to match the expected counts *exactly*? Why or why not?

	Season				Total
	Fall	Winter	Spring	Summer	
Observed Count	32	17	30	24	103

Note that we must allow for some slight variations in the crime patterns over the four seasons, no matter what the true probability of a crime occurring in any particular season is. Over repeated samples, slight variations will occur in the crime patterns simply because of chance.

Our goal is to determine whether the difference between the expected and observed counts is greater than we would expect to see by chance. We could use Tinkerplots to give us an idea of how much deviation from the expected counts we should anticipate due to randomness.
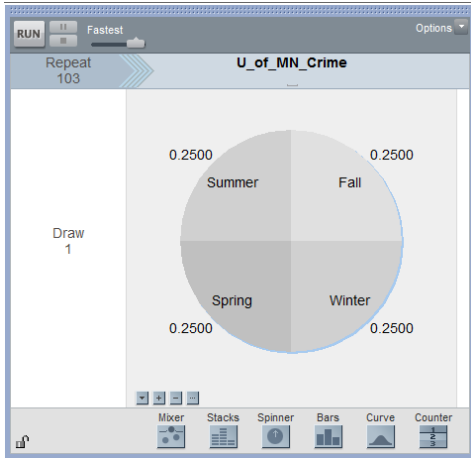
Questions:

1. Why is the spinner set up with 25% for each season? Explain.

2. Why is the repeat value set to 103? Explain.

Suppose that after running the first trial of the simulation, you obtained the following results:
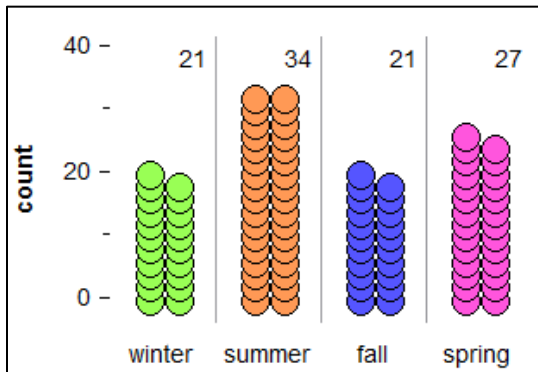

To keep track of the count for the number of crimes in each season over repeated trials of this simulation, we can right-click on *each* season's count and select **Collect Statistic**. Note that this needs to be repeated for each season. Tinkerplots then creates a table as follows:

	count_Attr1_winter	count_Attr1_summer	count_Attr1_fall	count_Attr1_spring	<new>
1	21	34	21	27	

History of Results of Sampler 1   Collect 1   Options

I placed 99 in the Collect box and then ran the simulation, so I ended with a total of 100 simulated trials.

	count_Attr1_winter	count_Attr1_summer	count_Attr1_fall	count_Attr1_spring	
97	21	26	22	34	
98	32	19	33	19	
99	20	28	27	28	
100	28	27	23	25	

History of Results of Sampler 1 — Collect 99 — \<new\> — Options

The following plot shows what outcomes our simulation study suggests we would anticipate for each season, assuming that crimes are equally likely to occur in any of the four seasons.
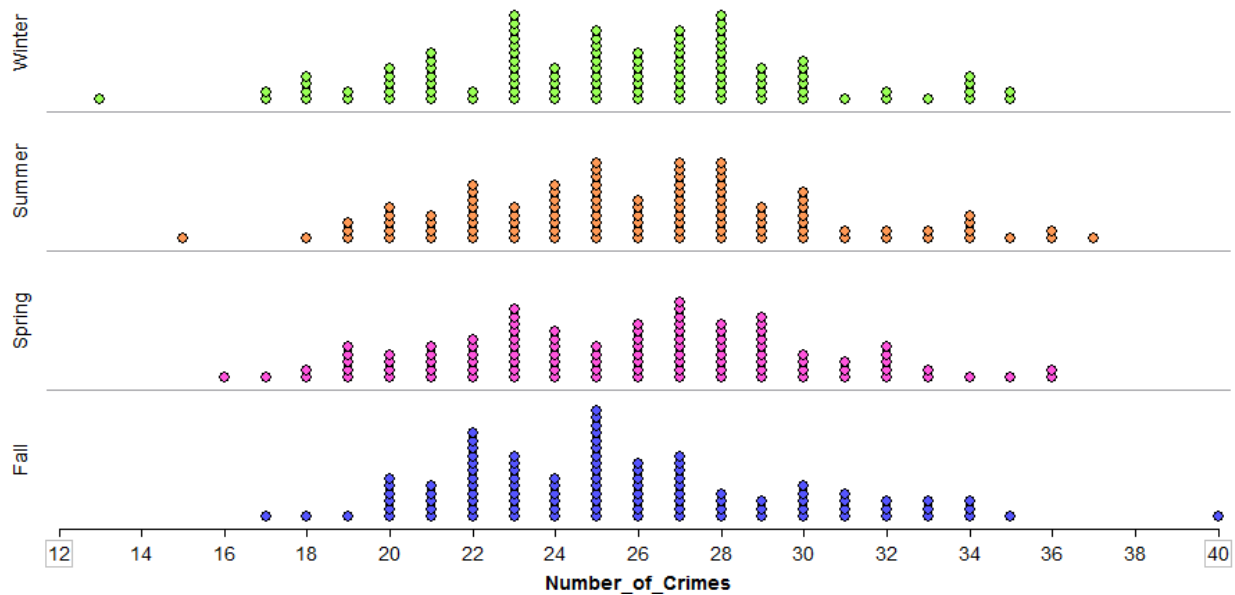
Next, recall the actual crime statistics for the University of Minnesota neighborhood for the past year.

	Season				Total
	Fall	Winter	Spring	Summer	
Observed Count	32	17	30	24	103

Plot these observed outcomes on the above graph. What do you think? Remember that our goal is to determine whether this observed result is inconsistent with what our simulation tells us to expect if there is really no difference in the crime rate across seasons.

Why is it more difficult to determine whether or not the observed data is considered an outlier in this situation than in the problems we dealt with previously?

Measuring Distance Between Observed and Expected Counts with Several Categories

One way to address this problem is to consider the overall distance between the observed and expected counts for all four seasons combined. This can be visualized as follows.
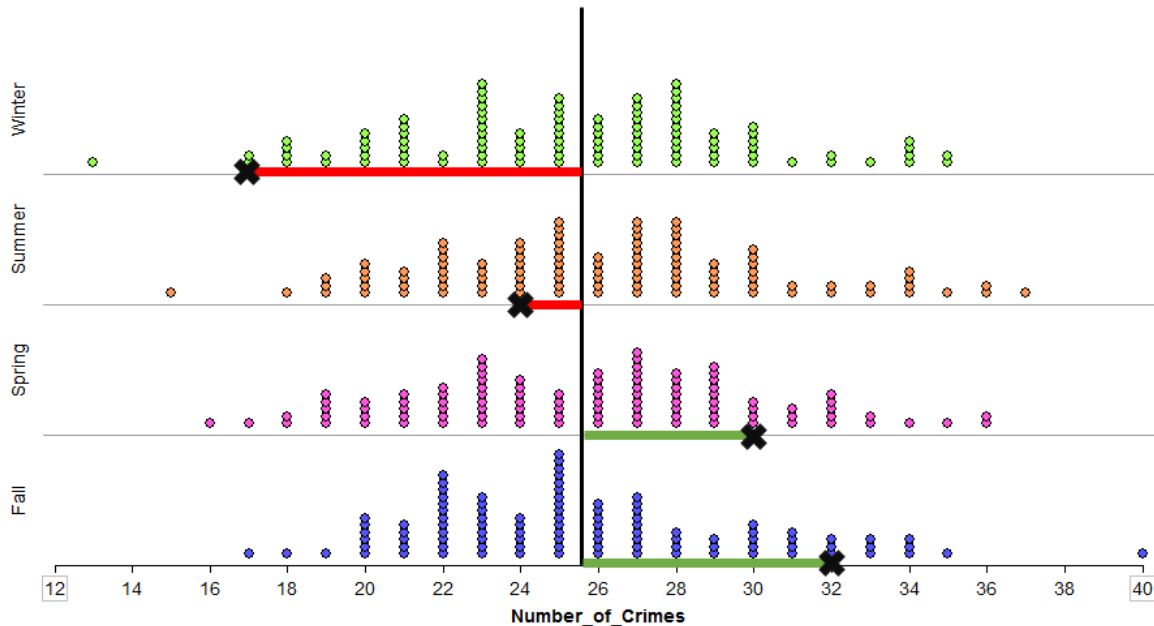
Compute the distance betweeen the observed and expected counts for each season in the table below.

	Season				Total
	Fall	Winter	Spring	Summer	
Observed	32	17	30	24	103
Expected	25.75	25.75	25.75	25.75	103
Distance					

Questions:

1. Find the sum of the distances between observed and expected counts over all four seasons. What is this total distance? Does this value make sense for total distance? How might we overcome this issue?

So that the negative distances do not cancel out the positive ones, we will square each distance before adding them up. *Note that the absolute value could have been used, as well, but the statistical hypothesis testing procedure we will soon be discussing uses the squared distances, so that's what we will consider here.*

	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.

Note that we cannot determine whether or not this 137 is an outlier using our previous graph of the simulated results because the previous graphs considered each season individually.

Our new measure is the squared distance between the observed and expected counts summed over four seasons, so we now need a new graph that shows this value for all of our 100 simulated results in Tinkerplots to determine whether or not 137 is an outlier.
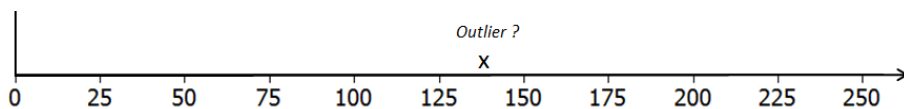
Questions:

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 an upper-tailed test. 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 count for each season. These squared distance values are then summed across the four seasons.

The outcomes from the first ten simulated trials are shown below.

	count_winter	count_summer	count_fall	count_spring	Winter_Dist_Sq	Summer_Dist_Sq	Fall_Dist_Sq	Spring_Dist_Sq	Sum
1	21	34	21	27	22.5625	68.0625	22.5625	1.5625	113.188
2	23	32	20	28	7.5625	39.0625	33.0625	5.0625	79.6875
3	18	26	26	33	60.0625	0.0625	0.0625	52.5625	60.1875
4	30	25	22	26	18.0625	0.5625	14.0625	0.0625	32.6875
5	21	30	29	23	22.5625	18.0625	10.5625	7.5625	51.1875
6	21	32	32	18	22.5625	39.0625	39.0625	60.0625	100.688
7	26	22	26	29	0.0625	14.0625	0.0625	10.5625	14.1875
8	23	28	28	24	7.5625	5.0625	5.0625	3.0625	17.6875
9	26	24	24	29	0.0625	3.0625	3.0625	10.5625	6.1875
10	23	23	32	25	7.5625	7.5625	39.0625	0.5625	54.1875

History of Results of Sampler 1

A graph of the sum of the squared distances from all 100 trials in Tinkerplots is given below. The p-value is determined using the proportion of dots greater than or equal to 137, the "observed outcome" from the study.
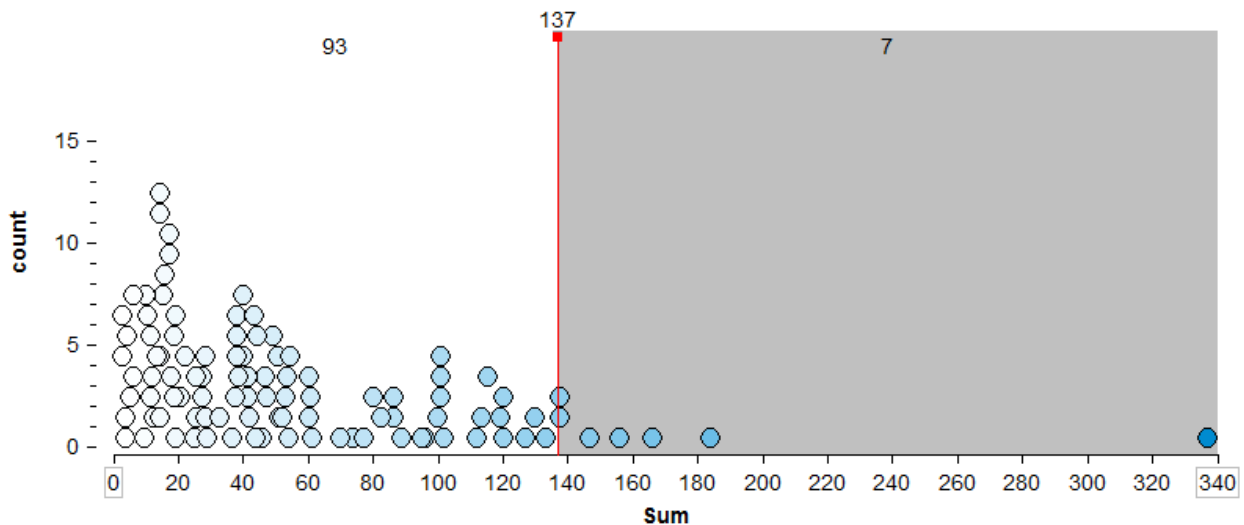
Questions:

1. What is the approximate p-value from the above graph?

2. What is the appropriate statistical decision for our research question?

Example 2.8: The Minnesota Student Survey (MSS) is a survey administered every three years to 6th, 9th, and 12th grade students. It is also offered to students in area learning centers and to youth in juvenile correctional facilities. This survey is an important vehicle for youth voice, as school districts, local public health agencies, and social services agencies use the survey results in planning and evaluation for school and community initiatives and prevention programming.

Questions are 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 this analysis, we will consider Question # 105 from this survey. Data has been collected for Fillmore County, which is in Southeastern Minnesota. Fillmore County (population = 20,866) consists of several small rural communities.

Question #105 from MN Students Survey

Fillmore County is in Southeastern Minnesota
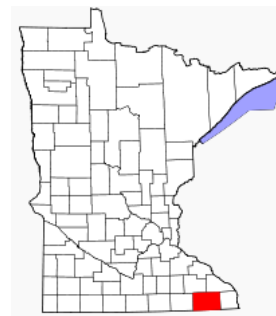
The following data was obtained from the Minnesota Department of Education website.

TABLE 39
PERCEIVED RISK OF HARM FROM SUBSTANCE USE
Fillmore County

			Grade					
			6th		9th		12th	
			Male	Female	Male	Female	Male	Female
			85	96	88	84	59	61
How much do you think people risk harming themselves physically or in other ways if they...	... smoke marijuana once or twice a week?	No risk	4	5	11	3	9	3
		Slight risk	3	6	1	1	8	8
		Moderate risk	10	10	16	13	14	15
		Great risk	68	75	60	67	28	35

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

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 Entire 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 statewide patterns of the marijuana portion of Question 105?
Parameters	The four parameters of interest are defined as follows: π_{no} = the probability a Grade 6 student from Fillmore County will respond to the marijuana portion of this question with No Risk π_{slight} = the probability a Grade 6 student from Fillmore County will respond to the marijuana portion of this question with Slight Risk $\pi_{moderate}$ = the probability a Grade 6 student from Fillmore County will respond to the marijuana portion of this question with Moderate Risk π_{great} = the probability a Grade 6 student from Fillmore County will respond to the marijuana portion of this question with Great Risk
Hypotheses	H_o: Fillmore County Grade 6 students do not deviate from historical statewide patterns H_a: Fillmore County Grade 6 students deviate from historical statewide patterns.

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). The second row contains the Expected 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:

3. How were the expected values computed?

4. Why is the *Expected* count for the Great Risk category so much higher than the others?

5. Suppose 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 from the historical percentages (i.e., No Risk = 1%, Slight Risk = 12%, Moderate Risk = 12%, and Great Risk = 75%)." Why is this statement statistically incorrect? Explain.

Once again, to understand the amount of acceptable deviation from the expected, we can consider the distance between the observed and expected counts 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 (O)	9	9	20	143	181
Expected (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 otherwise the positive and negative values cancel each other out. Upon careful inspection of these differences, there is another problem that also needs to be addressed.

Notice that the difference in the Great Risk category is 7.25, and the difference in the No Risk category is slightly smaller at 7.19. However, these distances alone fail to take into consideration the *scale* of the expected values. The following **Test Statistic** accounts for the scale of the expected counts:

$$Test\ Statistic = \sum \frac{(Observed - Expected)^2}{Expected}$$

	Survey Responses for Fillmore County, 6th Grade				Total
	No Risk	Slight Risk	Moderate Risk	Great Risk	
Observed (O)	9	9	20	143	181
Expected (E)	1.81	21.72	21.72	135.75	181
Difference (O - E)	7.19	-12.72	-1.72	7.25	
$(O - E)^2$	51.7	161.8	2.96	52.56	
$\frac{(O - E)^2}{E}$	28.56	7.45	0.14	0.39	36.5

The **Test Statistic** for this analysis would be:

$$\sum \left(\frac{(Observed - Expected)}{Expected} \right)^2 = \frac{(9 - 1.81)^2}{1.81} + \frac{(9 - 21.72)^2}{21.72} + \frac{(20 - 21.72)^2}{21.72} + \frac{(143 - 135.75)^2}{135.75}$$

$$= 28.56 + 7.45 + 0.14 + 0.39 = 36.5$$

Similar to what we have done in the past, now we must determine whether or not the Test Statistic from the observed data would be considered an outlier under the null hypothesis. Tinkerplots can be used to compute the Test Statistic over repeated samples under the null hypothesis.

The following graph shows the test statistic computed for each of the 100 simulations carried out under the null hypothesis in Tinkerplots:
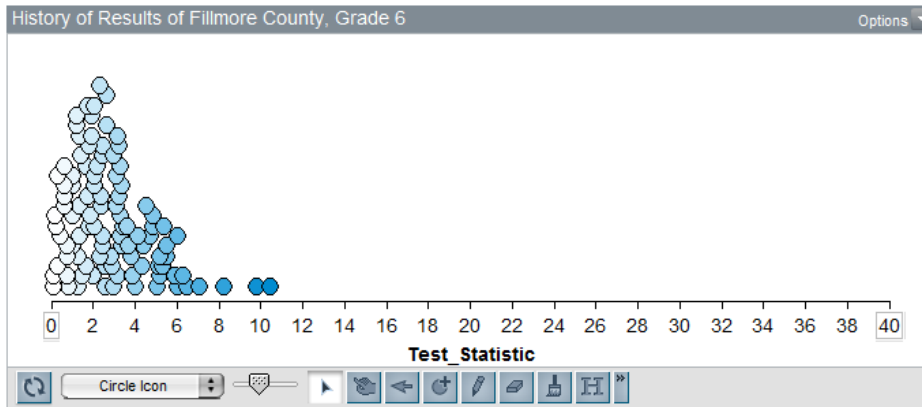
Questions:

1. The Test Statistic for the observed data from our sample of Grade 6 students in Fillmore County is 36.5. Is this value consistent with results we would expect to see if Grade 6 students in Fillmore County do not deviate from historical statewide patterns for the marijuana portion of this question? Explain.

2. Consider the formula for the test statistic. What would a value near 0 imply? What would a large value imply? Explain.

$$Test\ Statistic = \sum \frac{(Observed - Expected)^2}{Expected}$$

3. Use the simulated results to estimate the p-value, make a decision, and write a final conclusion in the context of the original research question.

The Chi-square Distribution

It can be shown that this test statistic actually follows what is known as the **Chi-Square Distribution**. This is different from the binomial distribution presented earlier.

- Unlike the binomial distribution, the chi-square distribution is not based on counts and is often skewed to the right.
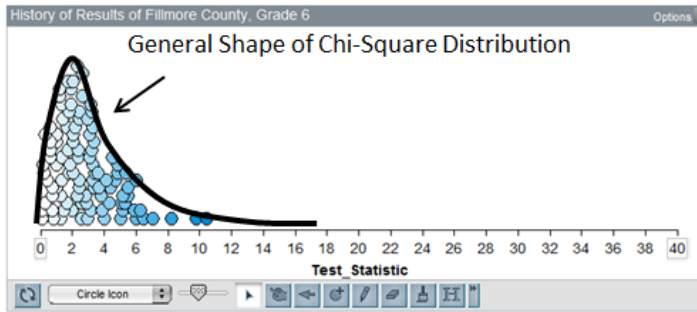
- The number of categories is taken into consideration through a quantity called the degrees of freedom, typically referred to as the **df**.

  df = # of Categories – 1

  The graph below shows several different chi-square distributions indexed by the degrees of freedom.
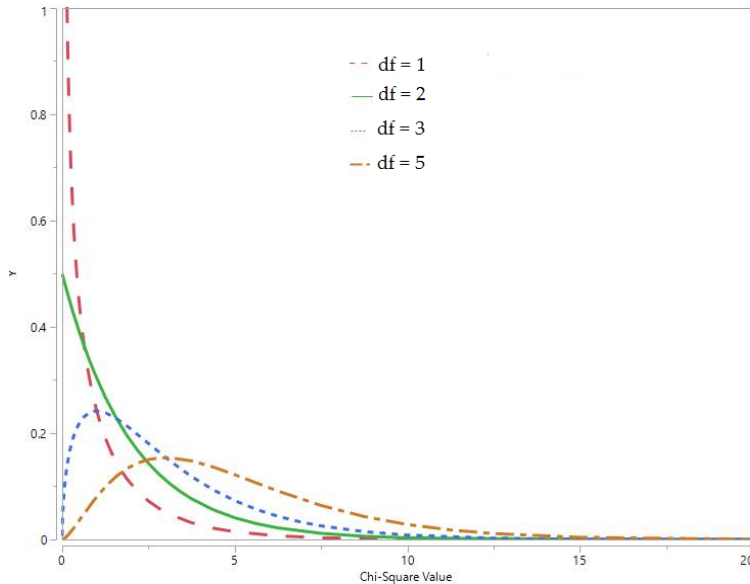
The Chi-Square Goodness of Fit Test

To test the hypothesis that Fillmore County sixth-graders deviate from historical statewide trends on this question, the researchers can conduct the test as follows:

- First, they set up their hypotheses.

  H_o: Fillmore County Grade 6 students do not deviate from historical statewide patterns
  H_a: Fillmore County Grade 6 students deviate from historical statewide patterns.

- Then, they try to get an understanding of what to expect if the null is true. Instead of using a simulation study to determine what to expect, we can use the chi-square distribution. The theory behind the chi-square test tells us that if the null hypothesis is true, the chi-square test statistic should follow a chi-square distribution with df = 3.

  Question: Why is df = 3 in this case?

  The chi-square distribution with df = 3 looks like this. This distribution helps us see what test statistics are likely (or unlikely) to occur if the null hypothesis is true.
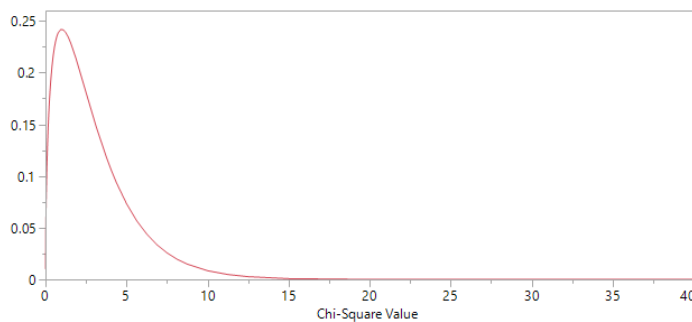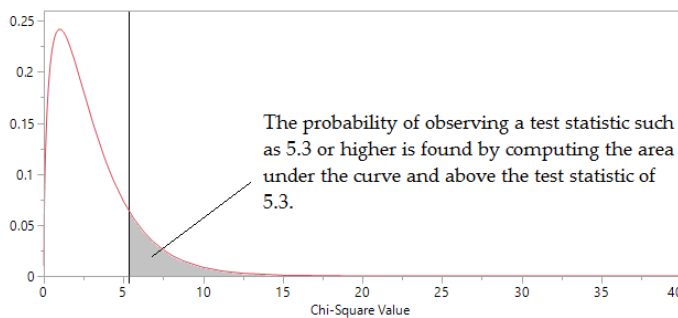

- Finally, they determine whether the test statistic computed from their data is consistent with what is expected under the null hypothesis. Recall that our test statistic was computed as follows:

$$\sum \left(\frac{\text{(Observed - Expected)}}{\text{Expected}} \right)^2 = \frac{(9-1.81)^2}{1.81} + \frac{(9-21.72)^2}{21.72} + \frac{(20-21.72)^2}{21.72} + \frac{(143-135.75)^2}{135.75}$$

$$= 28.56 + 7.45 + 0.14 + 0.39 = 36.5$$

  Clearly, a test statistic as large as 36.5 is very unlikely to happen by chance, so the p-value in this case will be *very* small. The survey results provide very strong evidence that Fillmore County Grade 6 students deviate from historical statewide patterns.

To find the p-value, we could use a chi-square probability calculator (like we did with the binomial distribution). To make life easier, however, we will simply let JMP calculate the p-value for us.

Carrying out the Chi-Square Goodness of Fit Test in JMP

This hypothesis testing procedure easily implemented in JMP when given the raw data. For example, consider the following data set:

Category	count
No Risk	9
Slight Risk	9
Moderate Risk	20
Great Risk	143

To get the test statistic and p-value in JMP, select **Analyze > Distribution**. Enter the following:
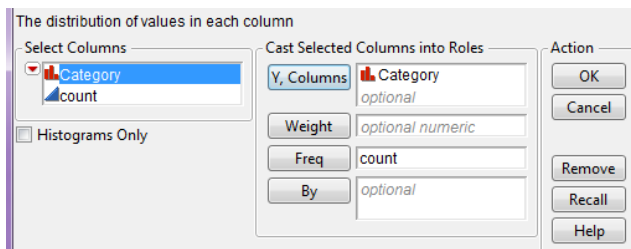
On the output that appears, select "Test Probabilities" from the red drop-down arrow:
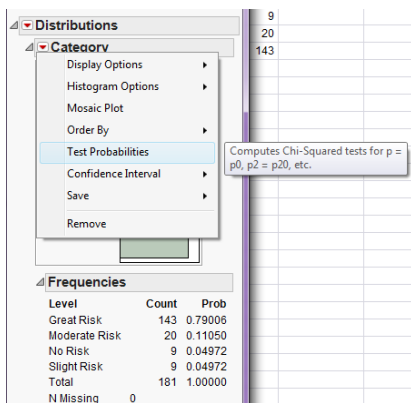
Enter the following:

Level	Estim Prob	Hypoth Prob
Great Risk	0.79006	0.75000
Moderate Risk	0.11050	0.12000
No Risk	0.04972	0.01000
Slight Risk	0.04972	0.12000

Click then Enter Hypothesized Probabilities.

Click "Done" and JMP should return this output:

Test Probabilities

Level	Estim Prob	Hypoth Prob
Great risk	0.79006	0.75000
Moderate risk	0.11050	0.12000
No risk	0.04972	0.01000
Slight risk	0.04972	0.12000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	24.5924	3	<.0001*
Pearson	36.5341	3	<.0001*

Note that the p-value associated with our chi-square test statistic (which was 36.5) is reported as <0.0001. *This means that the p-value is so small that it is less than 1 in 10,000.* The survey results provide very strong evidence that Fillmore County Grade 6 students deviate from historical statewide patterns.

Example 2.7 Revisited

To test the hypothesis that crimes were not occurring with equal probability over the four seasons, we can conduct the test as follows:

- First, we set up the hypotheses.

  H_o: Crimes are equally dispersed over the four seasons
  H_a: Crimes are not occurring with equal probability over the four seasons

- Then, we try to get an understanding of what to expect if the null is true. Recall that a chi-square distribution with 3 degrees of freedom tells us what values of the test statistic we expect to see when the null is true.

  Question: Why is df = 3 in this case?

  Here is the chi-square distribution with df =3.
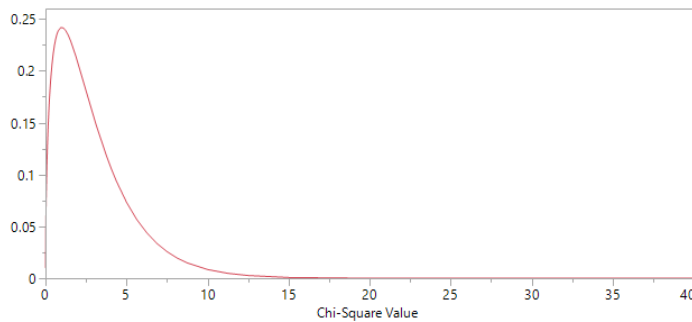
- Finally, we determine whether the test statistic computed from our data is consistent with what is expected under the null hypothesis. Our test statistic is computed as follows:

	Season				Total
	Fall	Winter	Spring	Summer	
Observed Count	32	17	30	24	103
Expected count	25.75	25.75	25.75	25.75	103

$$\sum \left(\frac{(\text{Observed - Expected})}{\text{Expected}} \right)^2 = \frac{(32 - 25.75)^2}{25.75} + \frac{(17 - 25.75)^2}{25.75} + \frac{(30 - 25.75)^2}{25.75} + \frac{(24 - 25.75)^2}{25.75} = 5.3$$

To find the p-value, we must determine how often we see test statistics such as 5.3 or higher by chance, even if the null is true. This probability can be visualized on the graph below (note that the total area under curve is 1.0).

The probability of observing a test statistic such as 5.3 or higher is found by computing the area under the curve and above the test statistic of 5.3.

Recall that we will use JMP to find this p-value.

Enter the raw data:	Use Analyze > Fit Distribution.
Test Probabilities:	The output:

What is our conclusion?