Chapter 3: Inference for a Single Categorical Variable with More than Two Categories

The analyses we completed in Chapter 2 were for a single categorical variable with only two outcomes. For example, in the helper/hinderer study, the babies were choosing either one toy or the other. In the sex discrimination example, each employee selected for management was either a woman or a man. Next, we'll consider problems involving a single categorical variable which has more than two categories.

Example 3.1: Fire Incidents

California is one of the places having the most deadliest and destructive wildfire seasons. This dataset contains the list of Wildfires that occurred in California between 2013 and 2020. The dataset contains the location where wildfires have occurred including the County name and also details on when the wildfire has started. A portion of the data set is shown below.

```
fire_incidents <- read_csv("data/fire_incidents.csv")                    ①
head(fire_incidents)                                                     ②
```

① Read the `fire_incidents` data set into the R environment, then
② print the first 6 rows of the data set.

```
# A tibble: 6 x 9
   Year Season Month Name       AcresBurned Counties CrewsInvolved Dozers Engines
  <dbl> <fct>  <chr> <chr>            <dbl> <chr>            <dbl>  <dbl>   <dbl>
1  2013 Summer Aug   Deer Fire        11429 Tehama              30      3      36
2  2013 Fall   Sep   Clover F~         8073 Shasta              12      3      30
3  2013 Summer Jul   Chariot ~         7055 San Die~            56     24     183
4  2013 Spring May   Panther ~         6965 Tehama              53     22     131
5  2013 Spring May   Grand Fi~         4346 Kern                29      1      34
6  2013 Fall   Nov   McCabe F~         3505 Sonoma               8     NA       5
```

This data is provided by kaggle.com (https://www.kaggle.com/datasets/ananthu017/california-wildfire-incidents-20132020/) to generate insights on what locations in California are under fire threat, what time of year do wildfires usually occur and how frequent and devastating they are. Suppose the fire chief in Riverside county has received the task of determining if the seasonal pattern of fires in the county differs from all of California.

? **Research Question:** Is there evidence to suggest that the proportion of fires occurring in each season in Riverside County is different from that of all of California?

Fires 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

Using the data, we can determine the proportion of fires occurring in each season across all of California. These proportions are given below:

```
fire_incidents |>                                   ①
  count(Season) |>                                  ②
  adorn_percentages(denominator = "all")            ③
```

① Start with the original `fire_incidents` data set (all of California), then
② count the number of total fires occurring in each season, finally
③ use the `adorn_percentages()` function to convert the counts to proportions (note: adds to about 1; aka 100%).

```
 Season  Prop
   Fall 0.219
 Winter 0.073
 Spring 0.112
 Summer 0.596
```

These are the Expected proportions based on ALL of California (the distribution we want to compare Riverside to).

The fire chief reported a total of 146 fires occurred in Riverside county between 2013 and 2020.

The code below shows how we can get only the fires in Riverside county from the original data set and save it to a data set named `riverside_incidents`

```
riverside_incidents <-                                                    ①
  fire_incidents |>                                                       ②
  filter(Counties == "Riverside")                                         ③

head(riverside_incidents)                                                 ④
```

① Create a new subset of data called `riverside_incidents`.
② Start with the original `fire_incidents` data set, then
③ filter the data set to include only Riverside county.
④ Preview the top 6 rows of the `riverside_incidents` data set.

```
# A tibble: 6 x 9
   Year Season Month Name      AcresBurned Counties CrewsInvolved Dozers Engines
  <dbl> <fct>  <chr> <chr>           <dbl> <chr>            <dbl>  <dbl>   <dbl>
1  2013 Summer Jul   Mountain~       27531 Riversi~            NA     NA      NA
2  2013 Summer Aug   Silver F~       20292 Riversi~            63     20     201
3  2013 Spring May   Summit F~        3166 Riversi~            36      5      62
4  2013 Summer Aug   Falls Fi~        1383 Riversi~            NA     NA      NA
5  2013 Spring May   Gorgonio~         510 Riversi~            NA     NA      NA
6  2013 Winter Feb   Jurupa F~         311 Riversi~            NA     NA      NA
```

Setup

Population All potential fires that could occur in Riverside.

Sample $n = 146$ fires which occurred in Riverside county between 2013 and 2020.

Observation A fire in Riverside county

Variable of interest Season (fall/winter/spring/summer)

Parameters of interest

The four parameters of interest are defined as follows:

π_{fall} = the population proportion of a Riversie fire occurring in the Fall
π_{winter} = the population proportion of a Riverside fire occurring in the Winter
π_{spring} = the population proportion of a Riverside fire occurring in the Spring
π_{summer} = the population proportion of a Riverside fire occurring in the Summer

Hypotheses

H_0: The proportion of fires occurring in each season in Riverside County is equal to the proportion in all of California.
($\pi_{\text{fall}} = 0.219$; $\pi_{\text{winter}} = 0.073$ $\pi_{\text{spring}} = 0.112$; $\pi_{\text{summer}} = 0.596$;
H_A: The proportion of fires occurring in each season in Riverside County is different from that in all of California (at least one seasonal proportion differs).
(at least one π_i differs).

The approach we take here to address the research question is very similar to what we have done previously. We will *assume* the the seasonal distribution of fires in Riverside County matches the seasonal distribution of fires across California (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 have sufficient statistical evidence to say the number of fires vary across the four seasons.

1. Find the expected number of fires in Riverside for each season under the assumption that fires are occurring with equal proportion to all of California. How did you obtain these values?

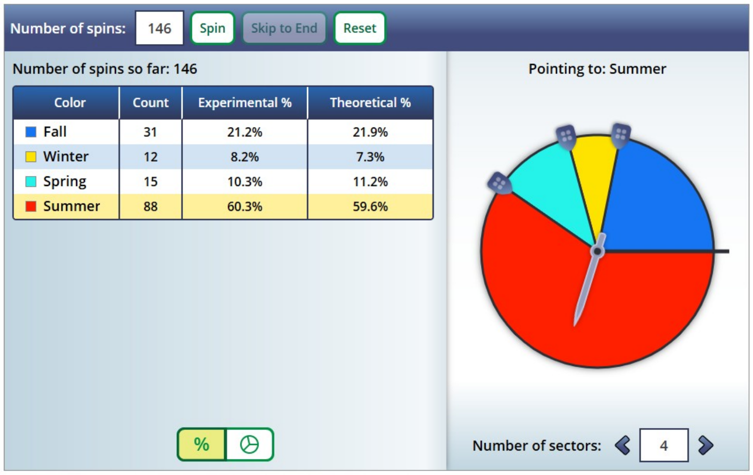
Expected Count = N x expected proportion

	Fall	Winter	Spring	Summer	Total
Expected Count	146(0.219) =31.97	146(0.073) =10.66	146(0.112) =16.35	146(0.596) =87.02	146

Note that we must allow for some slight variations in the fire patterns over the four seasons because we do not anticipate the numbers to come out exactly at the expected number for each season. Over repeated samples, slight variations will occur in the fire patterns. We could use simulation to give us an idea of how much deviation from the expected counts we should anticipate.

2. We will use an adjustable spinner to simulate the variations we would expect to see *under* the assumption that the seasonal distribution of fires in Riverside County matches the seasonal distribution of fires across California.

- Visit https://www.nctm.org/Classroom-Resources/Illuminations/Interactives/Adjustable-Spinner/.
- Change the `Number of sectors` to 4 and change the color names to the Season names.
- Click the edges of the spinner and drag to set equal to the proportions given for seasons in all of California. *Hint: Start by shrinking the blue region, then yellow, followed by the cyan, and finally enlarge the red area. Don't get frustrated getting exact.*



Number of spins: 146

Number of spins so far: 146

Color	Count	Experimental %	Theoretical %
■ Fall	31	21.2%	21.9%
■ Winter	12	8.2%	7.3%
■ Spring	15	10.3%	11.2%
■ Summer	88	60.3%	59.6%

Pointing to: Summer

Number of sectors: 4

Why should the spinner's proportions for each season be set to match the seasonal proportions of fires in all of California? Explain.

- Spin once (this is the equivalent of one coin toss). Spin a couple more times to see what happens to the count table.
- `Reset` the count.
- Set the `Number of Spins` to 146 and hit Spin (or Skip to End).

Why is do we want to set this to 146? Explain.

This is our sample size. We want to distribute the n = 146 fires in Riverside County to the four seasons *under the assumption the proportions follow the greater California distribution of fires.*

- Record the results of your simulation in the table below.

Your simulation results

Note this was my simulation – you might have different counts.

	Fall	Winter	Spring	Summer	Total
Simulated Counts	31	12	15	88	146

Next, consider the actual fire statistics for Riverside County between 2013 and 2020.

Observed data from Riverside County

```
riverside_incidents |>       ①
  count(Season)              ②
```

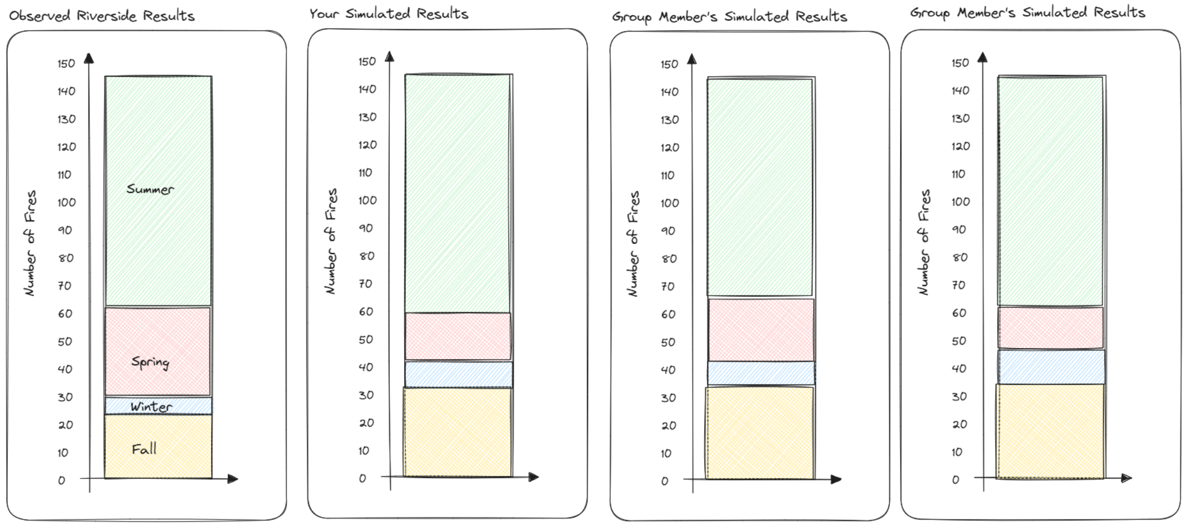
① Start with the `riverside_incidents` data set containing only fires in Riverside county, then

② summarize the data to obtain the number of fires in each `Season`.

```
# A tibble: 4 x 2
  Season     n
  <fct>  <int>
1 Fall      25
2 Spring    31
3 Summer    85
4 Winter     5
```

	Fall	Winter	Spring	Summer	Total
Observed Counts	25	5	31	85	146

3. Next to the stacked bar plot of the observed data, sketch out a stacked bar plot of your simulation results *under* the assumption that the seasonal distribution of fires in Riverside County matches the seasonal distribution of fires across California. Additionally sketch your other group members simulation results.



4. How does the observed Riverside data compare to the variations of simulated results *under* the assumption that the seasonal distribution of fires in Riverside County matches the seasonal distribution of fires across California?

The simulated counts (under the null) actually appear fairly similar to the expected counts making me think maybe the seasonal distribution of fires in Riverside County does match the seasonal distribution of fires across CA. Maybe there are a few more simulated fires in Fall than what we observed, but not super obvious. We will formally test this!

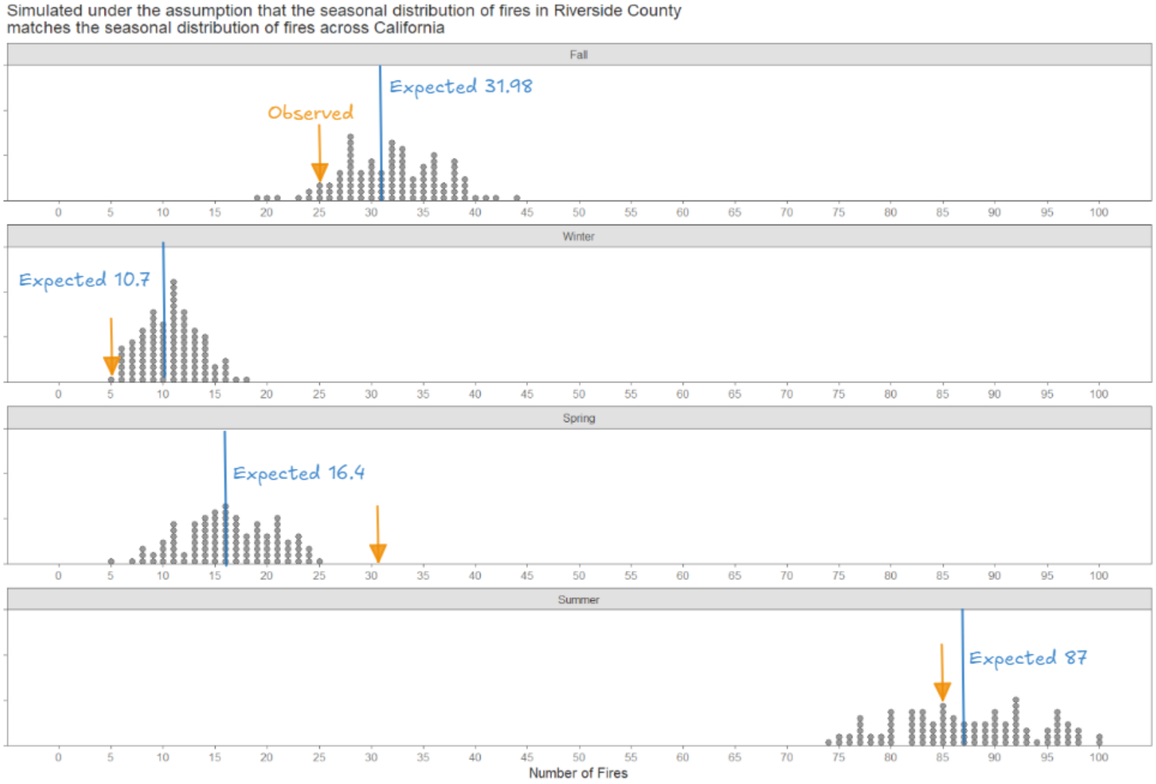
Recall that our goal is to determine whether this observed result is consistent or not with what our simulation tells us we should see if there really is no difference in the fire counts across seasons. To address the research question, we must identify whether or not the observed result would be considered "unusual" for each season. To formalize the process of determining whether our results are "unusual", recall that we can use the p-value approach. Recall the definition of a p-value.

Recall: p-value

The probability of observing an outcome as extreme or more extreme than the observed outcome *under* the assumption the null is true. This provides us a measure of the strength of evidence for the the research hypothesis (H_A).

Recall our **research question:** "Is there evidence to suggest that the proportion of fires occurring in each season in Riverside County is different from that of all of California?"

The following plot shows what outcomes our simulation study suggests we would anticipate for each season, *under* the assumption that the seasonal distribution of fires in Riverside County matches the seasonal distribution of fires across California.



5. Using the figure above, find the approximate *two-tailed* p-value for each season and make a decision.

Season	Obs. Count	p-value	Reject or FTR?
Fall	25	0.17	FTR
Winter	5	0.07	FTR
Spring	31	<0.0001	R
Summer	85	0.87	FTR

6. Why is it more difficult to determine whether or not our data is considered “unusual” in this situation than in the problems we dealt with previously?

There are multiple p-values (more than 2 categories) that lead to different conclusions.

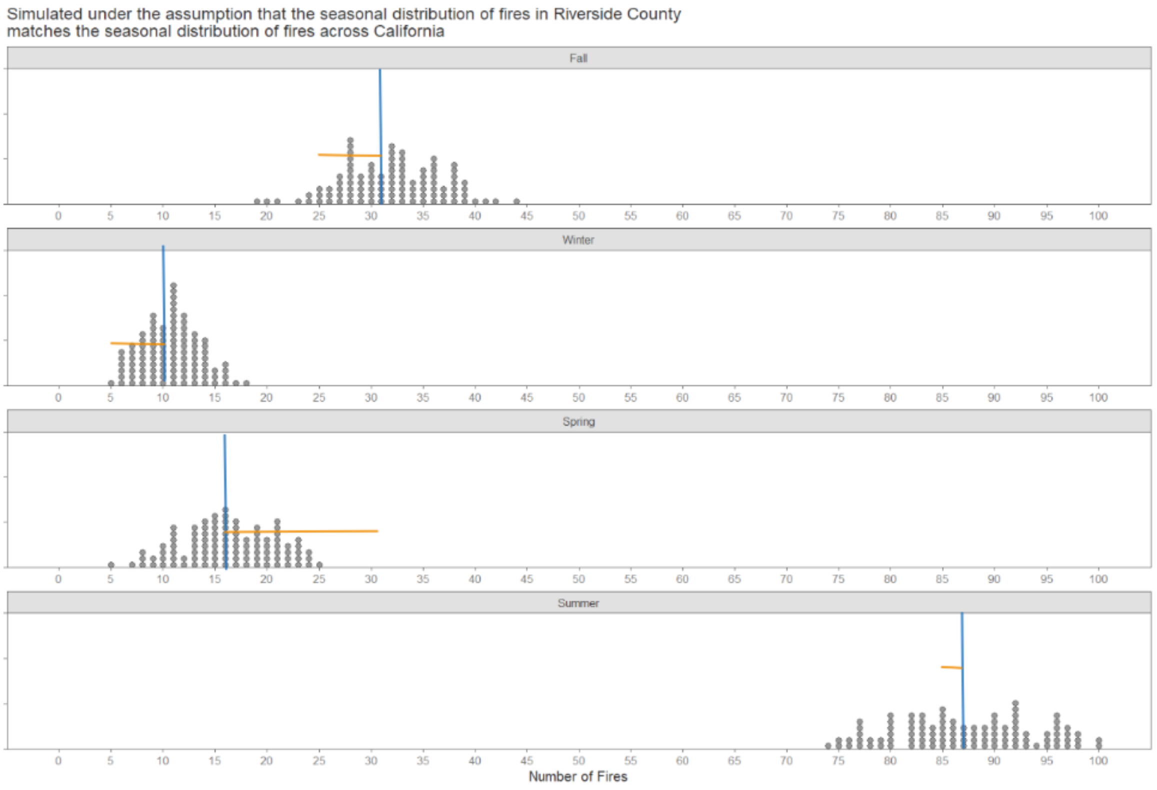
A comment regarding multiplicity

The issue of considering a set of p-values simultaneously to make a single decision is often referred to as multiplicity. Each of the above tests (one for each season) was conducted with an error rate of .05. One concern when using multiple p-values is that the “familywise” error rate ends up being much greater than .05, which is our gold standard for making decisions.

It can be shown that the maximum familywise error rate for a set of k tests, each conducted at the .05 error rate, is given by .05*k (this follows from Boole’s inequality). So, when we conduct four tests like we did above, our “familywise” error rate could be as large as .20!

We will talk more about this at the end of the quarter!

Measuring Distance between Observed and Expected with Several Categories As mentioned above, making a single decision with multiple p-values is problematic. One way to address this problem is to stop considering all four seasons separately. Instead, we could consider a single measure that combines information from all four seasons. For example, we could consider the overall distance between the Observed and Expected counts for all four seasons combined. This is shown below.



Total Distance ______ ____ ____________ __

Fall Winter Summer Winter

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

	Fall	Winter	Spring	Summer	Total
Observed	25	5	31	85	146
Expected	31.9	10.7	16.4	87	146
Distance	25 - 31.9 = -6.9	5 - 10.7 = -7.7			

8. Find the sum of the Distance for all four seasons. What is this total distance? Does this value make sense for total distance? How might we overcome this issue?

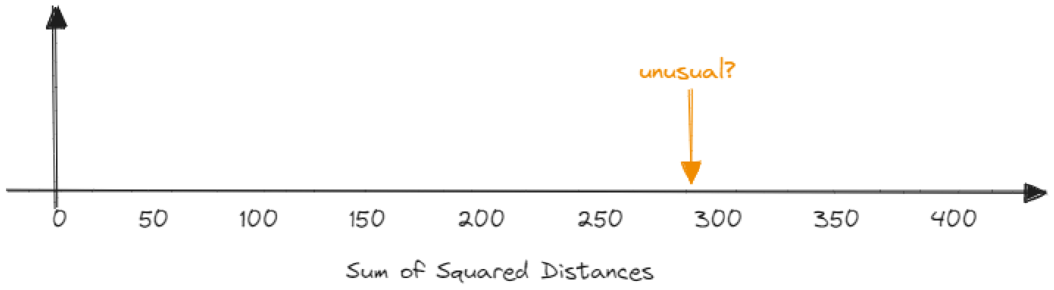
The sum of the distances is 0.. that doesn't help us. The positives and negative distances cancel out. We could take the absolute value? or square the distances. Let's square them!

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're considering here.

	Fall	Winter	Spring	Summer	Total
Observed	25	5	31	85	146
Expected	31.9	10.7	16.4	87	146
Distance	-6.9	-5.7	14.6	-2	0
Distance^2	47.61	32.49	213.16	4	**297.26**

The total squared distances summed up across all four seasons is about 297. Note that we cannot determine whether or not this 297 is "unusual" using our previous graphs because the previous graphs considered each season individually. Our new measure is the squared difference 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 from the online applet to determine whether or not 297 is "unusual".



9. What would a value of 0 imply on the above number line? Explain why this value less than 0 is not possible when the distances are squared and summed across the categories.

The expected counts are equal to the observed counts since squaring a negative makes it positive.

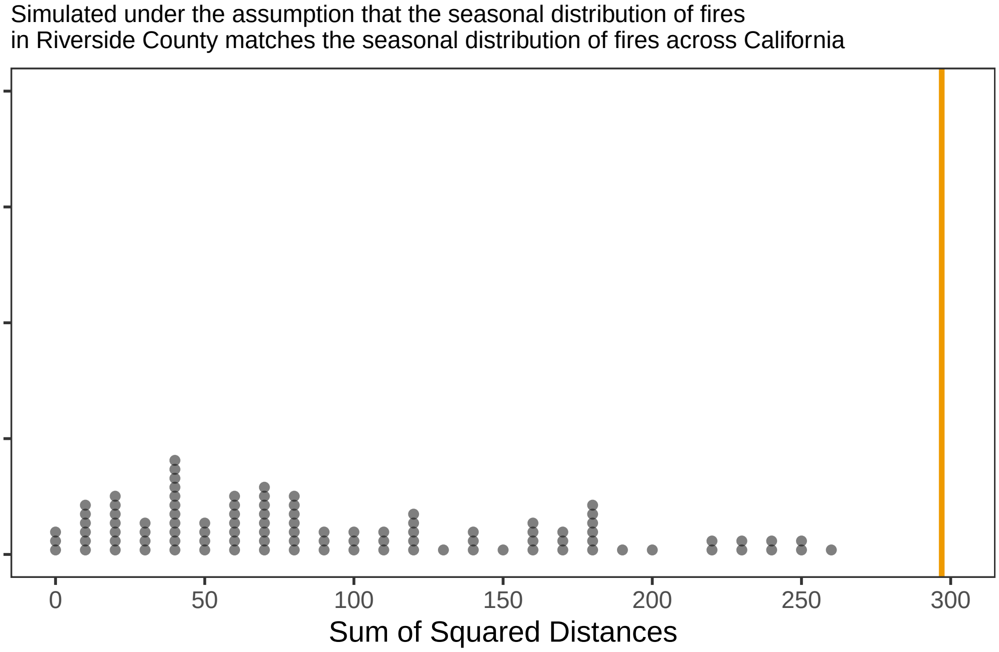
10. What would a large value imply? Is this evidence for or against the original research question?

The observed counts are VERY far from the expected counts. This is evidence for supporting our research question – if the distribution of fires in riverside is different from what we expect based on the distribution of CA, we would reject the null.

11. When squared distances are computed and summed across all categories, the appropriate test is an upper-tailed test. Explain why this is the case.

We squared the values so we already moved everything to one side.

A dot plot of the total squared distances from all 100 simulated sets of n = 146 fires under the assumption the is given below.



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

0/100 "dots" fall at or above our observed squared distance of 297.26. So a p-value = 0, but we say p-value < 0.0001.

However, the total of the squared distances fails under certain conditions since it does not take into consideration the *scale* of the expected values. The following **Test Statistic** accounts for the scale of the expected counts and is they typical summary measure used by statisticians.

ℹ Chi-square Test Statistic

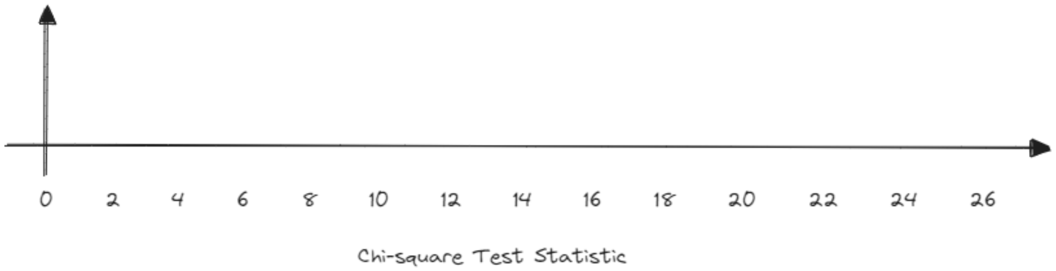
$$\text{Test Statistic} = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

	Fall	Winter	Spring	Summer	Total
Observed (O)	25	5	31	85	146
Expected (E)	31.9	10.7	16.4	87	146
Difference = (O - E)	-6.9	-5.7	14.6	-2	0
$(O-E)^2$	47.61	32.49	213.16	4	297.26
$\frac{(O-E)^2}{E}$	1.49	3.04			**17.98**

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

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 "unusual" under the null hypothesis.

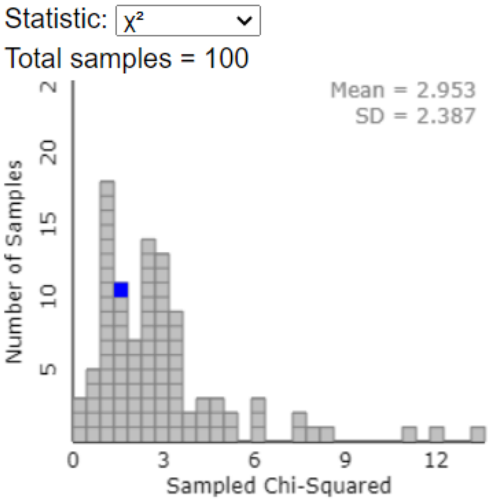
14. Return to your simulation from the spinner, and calculate the **Test Statistic** for the data simulated *under* the assumption that the seasonal distribution of fires in Riverside County matches the seasonal distribution of fires across California. Add a dot to the plot below. Now, add your group member's dots.



The `Online Simulation Applet > Goodness of Fit` can be used to compute the Test Statistic over repeated samples under the null hypothesis.

- Step 1: Clear the data, then Copy/Paste the data into the left window. Click `Use Data`
- Visit https:/raw.githubusercontent.com/earobinson95/stat218-calpoly/main/01-course-notes/data/riverside_seasons.csv to copy the `Season` variable/column of data
- Step 2: Make sure `Show Sampling Options` is selected.
- Step 3: Set *expected* probabilities – ⚠watch order of seasons! Click `Set`
- Step 4: Change `Number of Samples` to the number of repetitions you want to run (100 or 1000)
- Step 5: Change the `Statistic` above the simulated plot to X^2
- Step 6: Hit `Sample`
- Step 7: Similar to the one proportion applet, you can calculate/estimate your p-value by entering the observed Chi-square test statistic in the space below the plot.

The following graph shows the test statistic computed for 100 simulations carried out under the null hypothesis.



Statistic: X²

Total samples = 100

Mean = 2.953
SD = 2.387

15. The Test Statistic for the observed data from our 146 fires in Riverside County is $X^2 = 17.98$. Is this value consistent with results we would expect to see if the seasonal distribution of fires in Riverside County matches the seasonal distribution of fires across California? Explain.

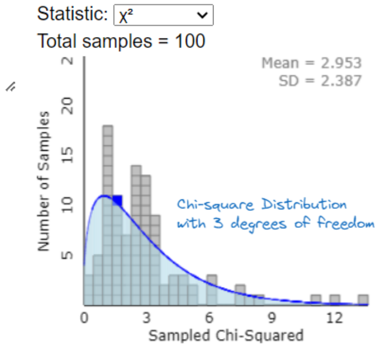
16. Consider the formula for the test statistic. What would a value near 0 imply? What would a large value imply? Explain.

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

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

The Chi-square Distribution and Goodness-of-fit Test

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 and Normal distribution, both presented earlier. Unlike the binomial distribution, the chi-square distribution is not based on counts and is often skewed to the right.



Statistic: χ²

Total samples = 100

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

$$\text{df} = \#\text{ of Categories}–1$$

We then refer to our Chi-square Test Statistic as χ^2_{df}.

This hypothesis testing procedure is often called the **Chi-square Goodness of Fit Test**. In practice, statisticians will use statistical software, such as R, to conduct a Chi-square Goodness-of-fit Test, given the data set (recall `riverside_incidents`).

The function `chisq_test()` from the `infer` package will conduct a Chi-square Goodness-of-fit-Test with the following parameter inputs:

- `x`: name of the data set
- `response`: name of the variable of interest in the data set
- `p`: a list of the categories and associated expected proportions – Caution! These need to be provided in alphabetical order.

```
library(infer)                                                      ①
chisq_test(x = riverside_incidents,                                 ②
           response = Season,                                       ③
           p = c("Fall"   = 0.219,
                 "Spring" = 0.112,
                 "Summer" = 0.596,
                 "Winter" = 0.073)                                  ④
           )                                                        ⑤
```

① Load the `infer` package.
② For the `chisq_test()` function, provide the data set input `x =`, and
③ indicate the variable of interest in `response =`.
④ Then provide the *expected* probabilities **in alphabetical order** for `p = c()`.
⑤ Ensure you close the function `()`

```
# A tibble: 1 x 3
  statistic chisq_df  p_value
      <dbl>    <dbl>    <dbl>
1      17.7        3 0.000509
```

19. State a conclusion for your research question.

⚠ Conditions for using the Chi-square Goodness of Fit Test

In order for the χ^2 distribution to be a good approximation of the true sampling distribution, we need to verify two conditions:

- The observations are independent
- We have a "large enough" sample size
  - This is checked by verifying there are at least 5 expected counts in each category

If the condition about expected cell counts is violated, we are forced to use the simulation-based method.

20. Check the conditions are met for the Chi-square distribution to be a good approximation of the true simulated sampling distribution.

- Independent observations?

- "Large enough" sample size?