# Chapter 11

## This starts with 11.7 Chi-Sq. Goodness of Fit Test

# Chi-Squared Goodness of Fit Test

## The general question:

Is the distribution of the values of a categorical variable consistent with what we expected?

I find this to be a very interesting sort of test because the test statistic is clearly “made up” in some sense.

It is supported by some theoretical statistics results, but that isn’t the point of this lesson.

# Example: Ice cream sales in October

- In my small grocery store, I sell three flavors of ice cream: vanilla, chocolate, and strawberry.
- Last October's data showed that the proportions of the flavors I sold were 0.30, 0.50, and 0.20 (in that order.)
- When next October is completed, I will be interested in testing the claim that
  - next October's proportions were consistent with the previous October's proportions
  - versus
  - the alternative that some of the proportions are not consistent with last year's claim.

# Hypotheses

$H_0: p_V = 0.30,\ p_C = 0.50,\ \text{and}\ \ p_S = 0.20$

$H_A:\ \text{Some}\ \ p_i\ \text{is not as specified in}\ \ H_0$

# Explore possible data

Here are three possible datasets for next October.
How well does each match last October?

| Dataset A | |
|---|---|
| Vanilla | 60 |
| Chocolate | 100 |
| Strawberry | 40 |
| Total | 200 |

| Dataset B | |
|---|---|
| Vanilla | 70 |
| Chocolate | 95 |
| Strawberry | 35 |
| Total | 200 |

| Dataset C | |
|---|---|
| Vanilla | 76 |
| Chocolate | 87 |
| Strawberry | 37 |
| Total | 200 |

# What did you do?

Stop here and actually DO something.

- Remember the claim.

$$H_O: p_V = 0.30,\ p_C = 0.50,\ \text{and}\ p_S = 0.20$$

$$H_A: \text{Some}\ p_i\ \text{is not as specified in}\ H_O$$

- Notice the total number of observations. (200)
- Get some numbers to compare to and go back to the data tables make the comparisons.

# How well did they match?

- Dataset A. The Ho proportions, converted to counts, exactly matched.
- Dataset B. They were fairly close. (It is pretty easy to compare these numbers to Dataset A to see how close.)
- Dataset C. Not very close at all. (They sold 16 more Vanilla, 13 fewer Chocolate, and 3 fewer Strawberry.)

# How can we measure “how close?”

- My initial thought was to look at “how many” different it was.
- But we probably want to take into account how large the actual numbers, so maybe “how many” needs to be somewhat modified.
  (Maybe as a proportion, so we see that the difference of 3 and 13 for Strawberry and Chocolate aren’t so different in proportions?)

# What will we actually use for this?

We'll use a proportion and then multiply it by something that gets us back to the same "units" (number of cartons sold) as the original data.

Remember it from the first line; understand it from the second line.

$$\frac{(\text{Observed}-\text{Expected})^2}{\text{Expected}}$$

$$=\frac{(\text{Observed}-\text{Expected})}{\text{Expected}}\cdot(\text{Observed}-\text{Expected})$$

# Do I have to understand why this statistic?

**No.** Most students of statistics don't.

Here's what a theoretical justification of it looks like.

- Learn to derive Generalized Likelihood Ratio tests. And derive a test statistic for a set of goodness-of-fit hypotheses.
- Learn a theoretical result about the asymptotic dist'n of a test statistic like this.
- Make a second order Taylor approximation of the result of the previous step. That results in our test statistic.

# Test statistic

Here's the formula for the test statistic.

The summation here is over all cells. So, in our problem there are three terms.

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

# What supports Ha?

Notice two things:

- The further away our data is from the “expected” values, the larger the numerators will be.
- All terms are positive.

So only values in the right tail support Ha.

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

# Test statistic

- The test statistic has a
  - chi-square dist'n
  - with the number of degrees of freedom equal to the number of "free cells"
  - when we assume the number of observations is given.
- For this problem, degrees of freedom is the number of cells minus 1.

# Footnote about “free cells”

- A common use of goodness-of-fit tests is to test whether the data fit a certain distribution.
- In that case, we would usually use the data to estimate the parameters of that distribution, so that would lower the number of “free cells” by the number of parameters we estimate.

# Actually doing the test

- We could compute all the expected values “by hand.” (Actually we did that when we were first looking at the three datasets.)
- Then we could then compute the test statistic “by hand.”
- But that’s tedious. Let’s use software.

# Does Dataset B fit our claim?

- On the next pages are StatKey output giving us a chi-squared value of 2.542.
- With df=3-1=2, we find the p-value (right-tailed area) of  0.281.
  We find a similar p-value from a randomization test.

This is not reasonable evidence of a difference from the expected values.

It is reasonable to assume that our data fit the proportions in Ho.

# Results Dataset B

**Detailed Sample Table**

| | Count |
|---|---|
| V | 70<br>60<br>1.667 |
| C | 95<br>100<br>0.25 |
| S | 35<br>40<br>0.625 |

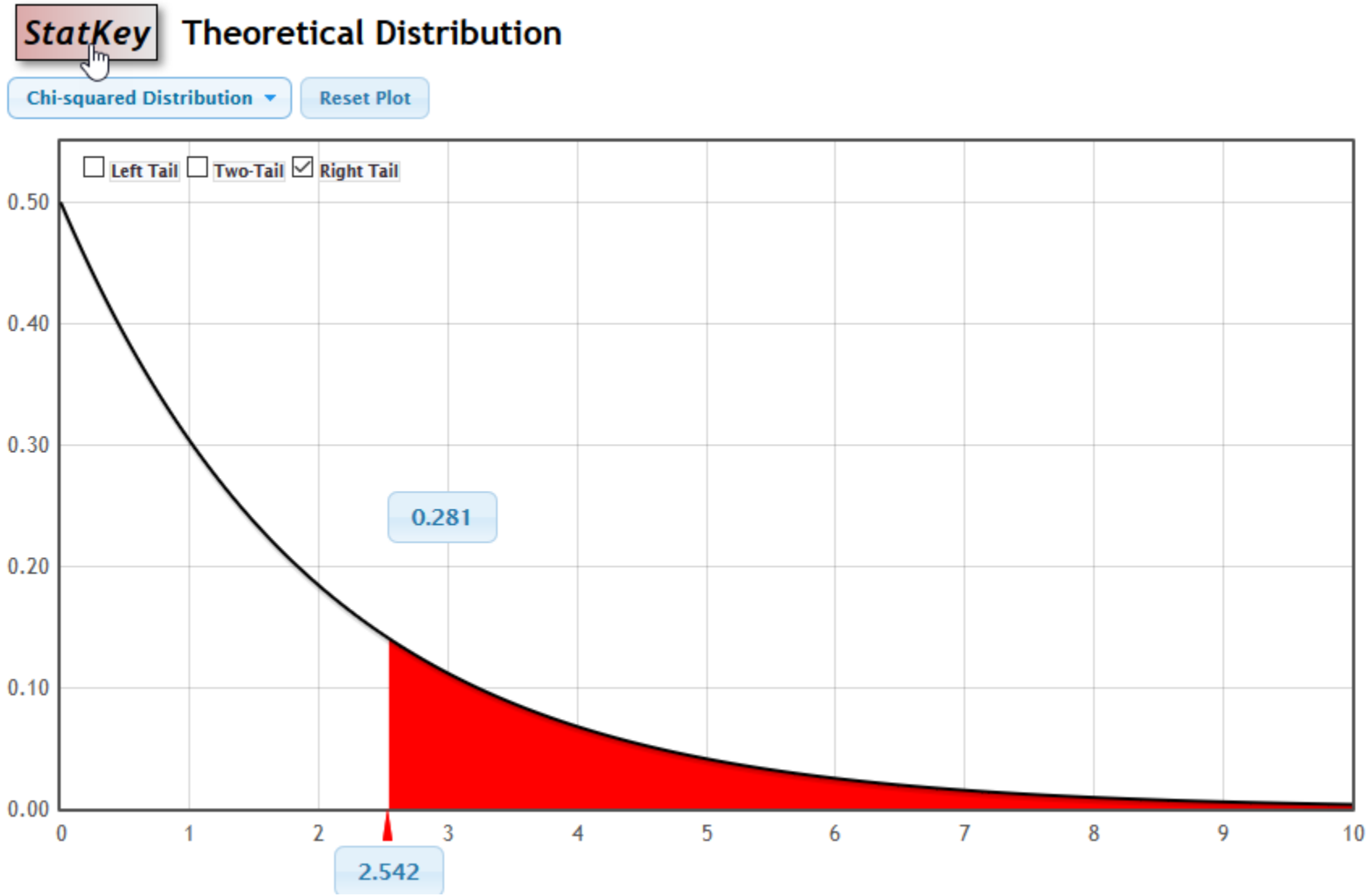
Observed, Expected, Contribution to $\chi^2$

**Original Sample** Show Details

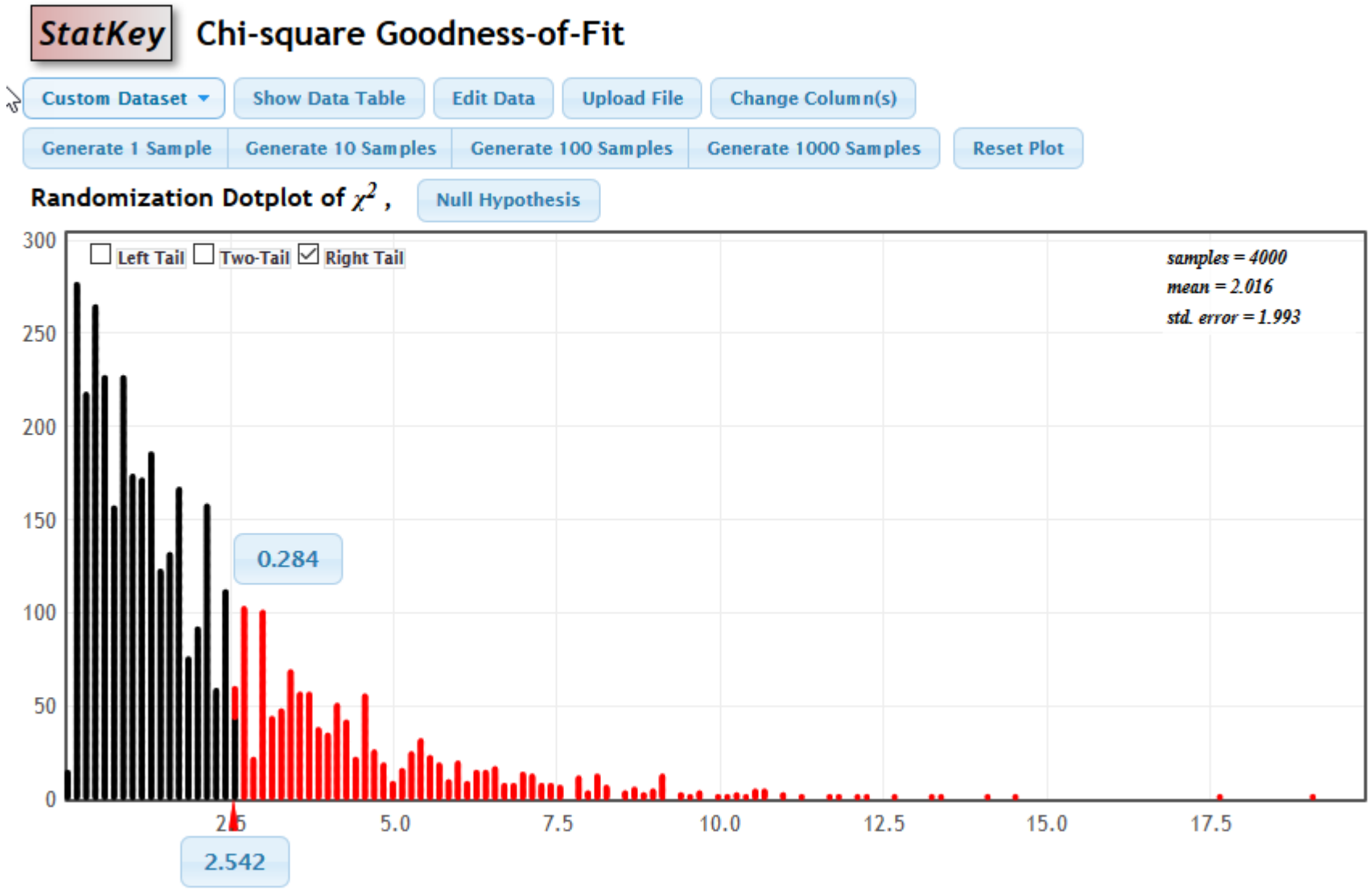
$n = 200, \chi^2 = 2.542$

| | Count |
|---|---|
| V | 70 |
| C | 95 |
| S | 35 |

# Dataset B: p-value from theoretical dist'n



# Dataset B: p-value from Randomization Dist'n



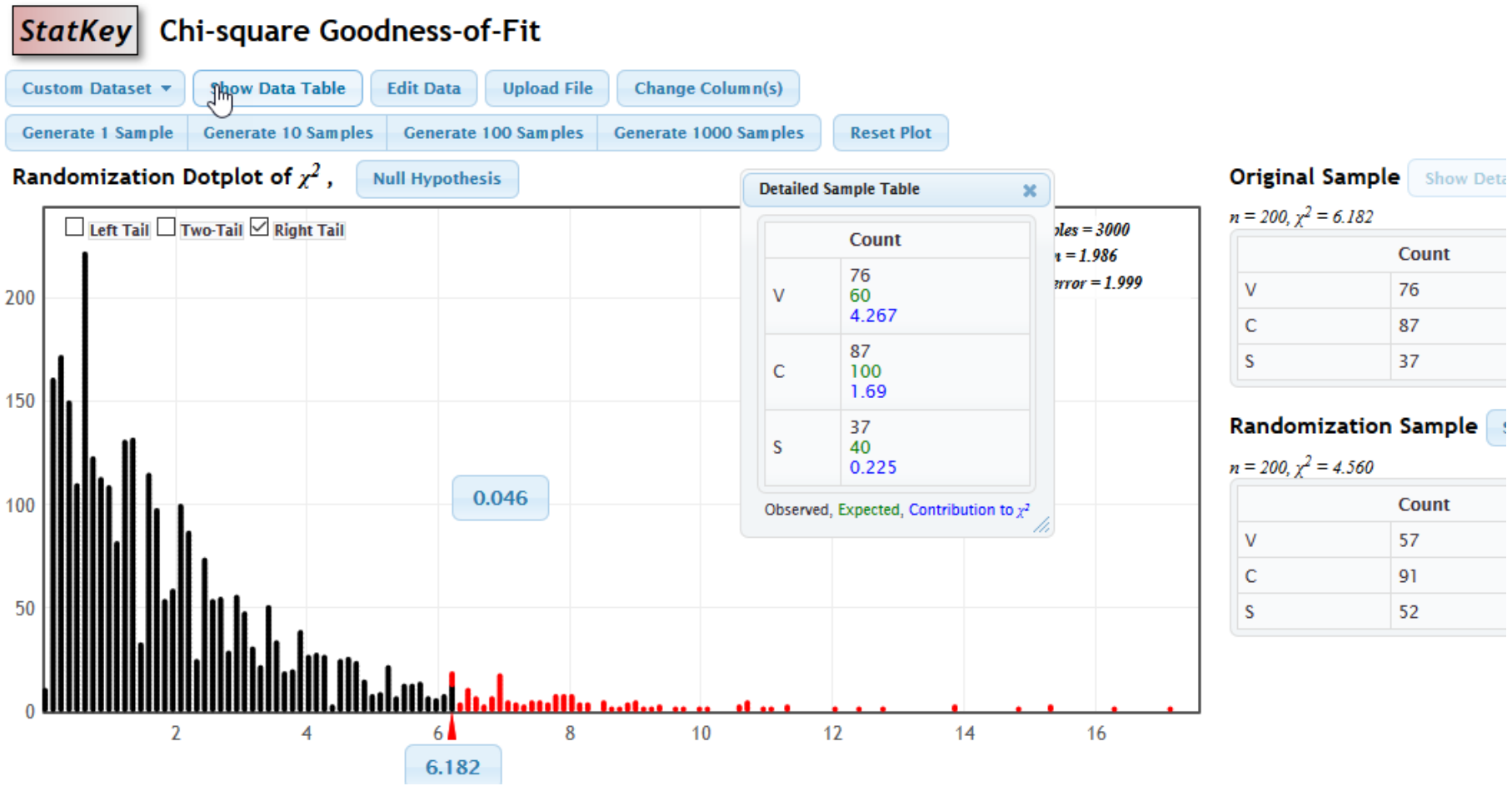
# Same question, but with Dataset C

Stop the video now and use your software (or work by hand) to practice the steps using Dataset C.

- Write hypotheses.
- Enter data. (Or compute the values needed and the chi-squared statistic.)
- Find the p-value (using either randomization or theoretical chi-squared dist'n.)
- Write a conclusion.

| Dataset C | |
|---|---|
| Vanilla | 76 |
| Chocolate | 87 |
| Strawberry | 37 |
| Total | 200 |

# Dataset C results



# Using **StatKey software**

It is important to BOTH

- EDIT Data

and

- Set the Ho values.

IN THAT ORDER.

(When you edit data, the Ho values revert to equal probabilities for all categories.)

# Conditions for Theoretical Dist'n

Of course, we should have checked the conditions for using the theoretical dist'n before we used it.

Here it is:

"Each expected count is at least 5."

That was met for this claim and these data.

# Goodness-of-fit footnote about conditions

- For goodness-of-fit tests, it is sometimes the case that you have some flexibility in what you decide goes into different categories of the data.
- 
- For this ice cream example, I might have started with six ice cream flavors – some of which sold very little. If I wanted to do a goodness-of-fit test, I could have combined some or all of the flavors with low sales into "Other" in order to make that category have a large enough expected value to use a chi-squared statistic.

# Comment about teaching these tests

- I have found (and heard from other teachers) that sometimes, even often, students say that, after learning this method, they “get it” about hypothesis testing, even when it was hard earlier.
- I wonder whether you feel that way.
- I wonder whether it’s because, for this claim, we usually “want” the Ho to be true. Maybe that’s what makes it feel simpler.
- I’d welcome your thoughts, if you have any on this.

# Chapter C

End of section

Next: C.8 Test of Association

# Chi-Squared Test of Association

The general question:

Is there a relationship between two categorical variables?

# Example:  Ice cream sales: Two locations

- In my small grocery stores, I sell three flavors of ice cream:  vanilla, chocolate, and strawberry.
- Is there a relationship between which store (location) and the pattern of ice cream flavors sold?

# Ways to state the hypotheses

Ho:  Location and the flavors of ice cream sold are not related.
Ha:  Location and the flavors of ice cream sold are related.

Ho: Location and the flavors of ice cream are independent.
Ha: Location and the flavors of ice cream are not independent.

Ho:  Location is not associated with flavors of ice cream sold.
Ha:  Location is associated with flavors of ice cream sold.

# What does “a relationship” mean?

- For the following possible observations about the situation, identify whether each supports the claim that there **is not** a relationship (Ho) or that there **is** a relationship (Ha.)

1. In the East location a higher proportion of chocolate ice cream is sold than in the West location.

Choose one: Supports Ho | Supports Ha | Not relevant to the claims

2. The same proportion of strawberry ice cream is sold in both locations.

Choose one: Supports Ho | Supports Ha | Not relevant to the claims

# Answers: What does “a relationship” mean?

- For the following possible observations about the situation, identify whether each supports the claim that there **is not** a relationship (Ho) or that there **is** a relationship (Ha.)

1. In the East location a higher proportion of chocolate ice cream is sold than in the West location.

Choose one: Supports Ho | **Supports Ha** | Not relevant to the claims

2. The same proportion of strawberry ice cream is sold in both locations.

Choose one: **Supports Ho** | Supports Ha | Not relevant to the claims

# Look at some data: 1

**Example (continued):** Here are three possible data sets for next October. Each value is the number of cartons of that flavor sold in October.

**Comment:** If there were just two categories for EACH variable, we could test it with a test of two proportions. But **we need something else to test whether multiple proportions fit a claim.**

| Dataset A | | | |
|---|---|---|---|
| | East | West | Total |
| Vanilla | 60 | 30 | 90 |
| Chocolate | 100 | 50 | 150 |
| Strawberry | 40 | 20 | 60 |
| Total | 200 | 100 | 300 |

p-value < 0.10? Yes Maybe No

| Dataset B | | | |
|---|---|---|---|
| | East | West | Total |
| Vanilla | 61 | 29 | 90 |
| Chocolate | 98 | 52 | 150 |
| Strawberry | 41 | 19 | 60 |
| Total | 200 | 100 | 300 |

p-value < 0.10? Yes Maybe No

| Dataset C | | | |
|---|---|---|---|
| | East | West | Total |
| Vanilla | 51 | 39 | 90 |
| Chocolate | 107 | 43 | 150 |
| Strawberry | 42 | 18 | 60 |
| Total | 200 | 100 | 300 |

p-value < 0.10? Yes Maybe No

# Look at some data: 2

Just by looking at them, think about what your conclusion to the hypothesis test is likely to be for each possible dataset. (Guess whether each is likely to have a p-value under 10%.)

Dataset A

| | East | West | Total |
|---|---|---|---|
| Vanilla | 60 | 30 | 90 |
| Chocolate | 100 | 50 | 150 |
| Strawberry | 40 | 20 | 60 |
| Total | 200 | 100 | 300 |

p-value < 0.10? Yes Maybe No

Dataset B

| | East | West | Total |
|---|---|---|---|
| Vanilla | 61 | 29 | 90 |
| Chocolate | 98 | 52 | 150 |
| Strawberry | 41 | 19 | 60 |
| Total | 200 | 100 | 300 |

p-value < 0.10? Yes Maybe No

Dataset C

| | East | West | Total |
|---|---|---|---|
| Vanilla | 51 | 39 | 90 |
| Chocolate | 107 | 43 | 150 |
| Strawberry | 42 | 18 | 60 |
| Total | 200 | 100 | 300 |

p-value < 0.10? Yes Maybe No

Look at Dataset A.

What are the proportions of the three flavors in the East location?
V: ____ C: _____ S:_____

What are the proportions of the three flavors in the West location?
V: ____ C: _____ S:_____

Explain why the proportions being equal for the two locations is evidence for the Ho.

# Look at some data: 2 (Answers)

Just by looking at them, think about what your conclusion to the hypothesis test is likely to be for each possible dataset. (Guess whether each is likely to have a p-value under 10%.)

| Dataset A | | | |
|---|---|---|---|
| | East | West | Total |
| Vanilla | 60 | 30 | 90 |
| Chocolate | 100 | 50 | 150 |
| Strawberry | 40 | 20 | 60 |
| Total | 200 | 100 | 300 |

p-value < 0.10? Yes Maybe No

| Dataset B | | | |
|---|---|---|---|
| | East | West | Total |
| Vanilla | 61 | 29 | 90 |
| Chocolate | 98 | 52 | 150 |
| Strawberry | 41 | 19 | 60 |
| Total | 200 | 100 | 300 |

p-value < 0.10? Yes Maybe No

| Dataset C | | | |
|---|---|---|---|
| | East | West | Total |
| Vanilla | 51 | 39 | 90 |
| Chocolate | 107 | 43 | 150 |
| Strawberry | 42 | 18 | 60 |
| Total | 200 | 100 | 300 |

p-value < 0.10? Yes Maybe No

Look at Dataset A.

What are the proportions of the three flavors in the East location?
V: 60/200 = 0.30 C: 100/200 = 0.50 S: 40/200 =0.20

What are the proportions of the three flavors in the West location?
V: 30/100 = 0.30 C: 50/100 = 0.50 S: 20/100 =0.20

Explain why the proportions being equal for the two locations is evidence for the Ho.

# Look at some data: 3

| Dataset A | | | |
|---|---|---|---|
| | East | West | Total |
| Vanilla | 60 | 30 | 90 |
| Chocolate | 100 | 50 | 150 |
| Strawberry | 40 | 20 | 60 |
| Total | 200 | 100 | 300 |

p-value < 0.10? Yes Maybe No

| Dataset B | | | |
|---|---|---|---|
| | East | West | Total |
| Vanilla | 61 | 29 | 90 |
| Chocolate | 98 | 52 | 150 |
| Strawberry | 41 | 19 | 60 |
| Total | 200 | 100 | 300 |

p-value < 0.10? Yes Maybe No

| Dataset C | | | |
|---|---|---|---|
| | East | West | Total |
| Vanilla | 51 | 39 | 90 |
| Chocolate | 107 | 43 | 150 |
| Strawberry | 42 | 18 | 60 |
| Total | 200 | 100 | 300 |

p-value < 0.10? Yes Maybe No

Look at Datasets B and C.

Do both show some difference in the distributions of flavors at the two stores? Yes

Which shows a stronger difference? C How can you tell? its values are further from what we see in A

Which would you expect to have a smaller p-value when you test the hypotheses? C

# Test Statistic

We will use the same chi-squared test statistic for this as we did for the goodness-of-fit test, where we compare the observed count in each cell with the expected count if Ho is true.

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

How do we find the degrees of freedom?

How do we find the expected counts?

# Degrees of Freedom

Given all the totals, the way the numbers are filled in describes the association (or lack of it.) The degrees of freedom is the number of blank cells below that I can FREELY fill in, while still keeping a table with these totals.

| | East | West | Total |
|---|---|---|---|
| Vanilla | | | 90 |
| Chocolate | | | 150 |
| Strawberry | | | 60 |
| Total | 200 | 100 | 300 |

# Degrees of Freedom

Stop the video here and fill in a number for any blank cell, and then see what you can still fill in and make sure

| | East | West | Total |
|---|---|---|---|
| Vanilla | | | 90 |
| Chocolate | | | 150 |
| Strawberry | | | 60 |
| Total | 200 | 100 | 300 |

# First try

Here's what I tried first, and it didn't work, because I couldn't make the totals work.

| | East | West | Total |
|---|---|---|---|
| Vanilla | 5 | | 90 |
| Chocolate | 5 | | 150 |
| Strawberry | | | 60 |
| Total | 200 | 100 | 300 |

| | East | West | Total |
|---|---|---|---|
| Vanilla | 5 | 85 | 90 |
| Chocolate | 5 | 145 | 150 |
| Strawberry | 190 | | 60 |
| Total | 200 | 100 | 300 |

# After several tries

After several more tries (starting with the third row) I found some numbers I could make work.

| | East | West | Total |
|---|---|---|---|
| Vanilla | | | 90 |
| Chocolate | 80 | | 150 |
| Strawberry | 50 | | 60 |
| Total | 200 | 100 | 300 |

| | East | West | Total |
|---|---|---|---|
| Vanilla | 70 | 20 | 90 |
| Chocolate | 80 | 70 | 150 |
| Strawberry | 50 | 10 | 60 |
| Total | 200 | 100 | 300 |

# Degrees of freedom

The purpose of all of this discussion is to clarify that there are only two “free” cells here.

r = number of rows,  c = number of columns

The formula for the degrees of freedom is

$df = (r - 1)(c - 1)$  (Here: df=(3-1)(2-1) = (2)(1) = 2)

# Expected Counts

In order to use the chi-squared statistic, we must have “expected” counts if Ho is true.

# Expected counts if Ho true 1

If the two variables are not associated, then we expect the same proportions as the total has in both colu

| | East | West | Total |
|---|---|---|---|
| Vanilla | $\frac{90}{300} \cdot \text{total}$ | $\frac{90}{300} \cdot \text{total}$ | 90 |
| Chocolate | $\frac{150}{300} \cdot \text{total}$ | $\frac{150}{300} \cdot \text{total}$ | 150 |
| Strawberry | $\frac{60}{300} \cdot \text{total}$ | $\frac{60}{300} \cdot \text{total}$ | 60 |
| Total | 200 | 100 | 300 |

# Expected counts if Ho true 2

The "total" in this diagram is the column total in each case.

| | East | West | Total |
|---|---|---|---|
| Vanilla | $\frac{90}{300} \cdot 200$ | $\frac{90}{300} \cdot 100$ | 90 |
| Chocolate | $\frac{150}{300} \cdot 200$ | $\frac{150}{300} \cdot 100$ | 150 |
| Strawberry | $\frac{60}{300} \cdot 200$ | $\frac{60}{300} \cdot 100$ | 60 |
| Total | 200 | 100 | 300 |

# Formula for Expected Counts

Typically, the formula for expected counts is communicated as

$$\frac{\text{Row Total} \cdot \text{Column Total}}{\text{Sample Size}}$$

Check to see that this is equivalent to what we found from our proportion discussion on the previous two slides.

# Computing the chi-squared statistic

Generally speaking, we don't compute this chi-squared statistic by hand – there are too many calculations.

As we learned with the previous chi-squared test, it is the large values of the statistic that are evidence against Ho so the rejection region is

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

# Testing Dataset C in StatKey

Use the chi-squared Test for Association

Put the data in using Edit Data and follow the format of the data that you see there.

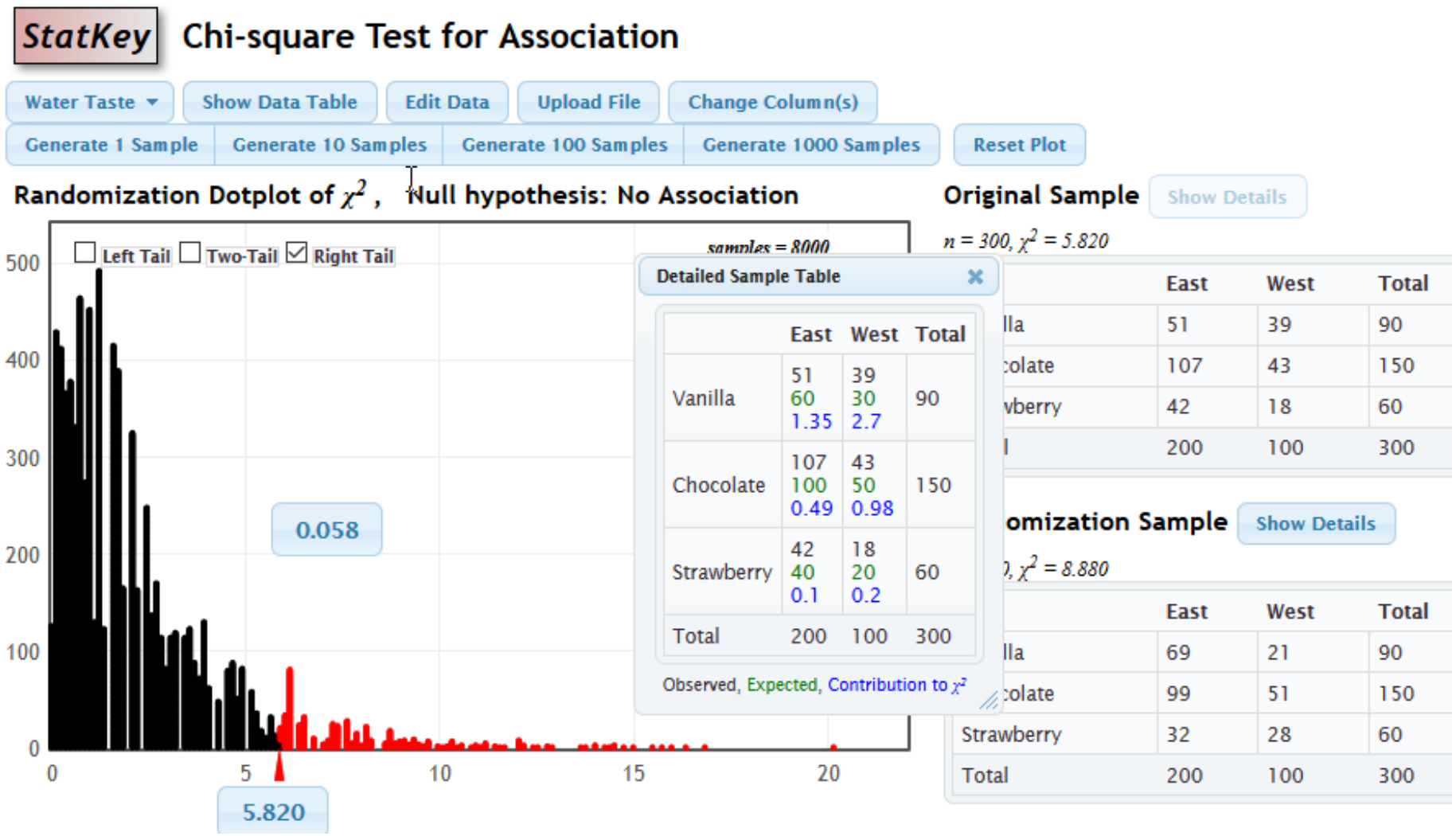
[blank],East,West

Vanilla,51,39

Chocolate,107,43

Strawberry,42,18

# Dataset C Randomization test output




**Original Sample** Show Details

$n = 300, \chi^2 = 5.820$

| | East | West | Total |
|---|---|---|---|
| lla | 51 | 39 | 90 |
| colate | 107 | 43 | 150 |
| vberry | 42 | 18 | 60 |
| l | 200 | 100 | 300 |

**omization Sample** Show Details

$, \chi^2 = 8.880$

| | East | West | Total |
|---|---|---|---|
| lla | 69 | 21 | 90 |
| colate | 99 | 51 | 150 |
| Strawberry | 32 | 28 | 60 |
| Total | 200 | 100 | 300 |

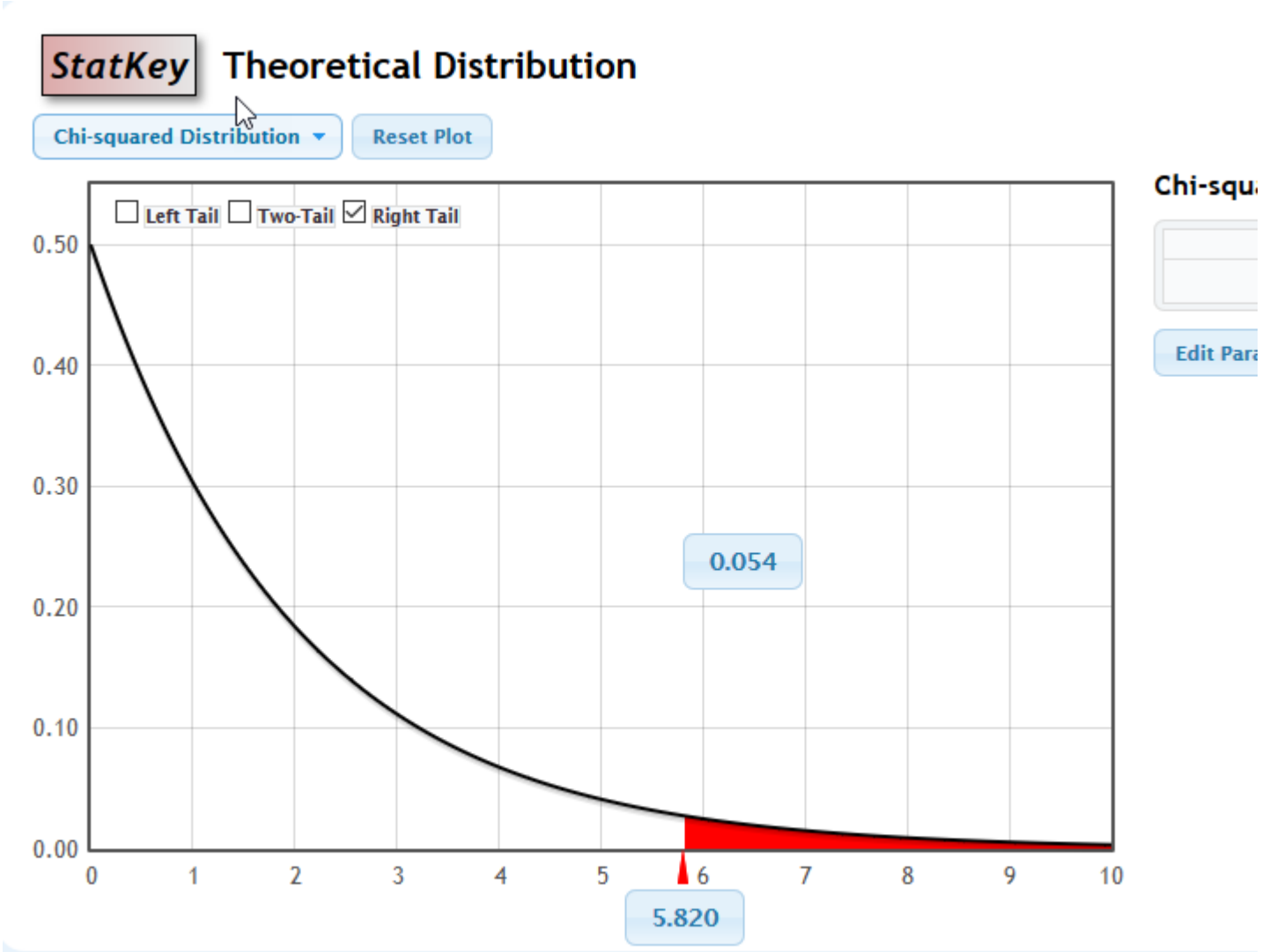
# Conditions to use the theoretical dist'n

The conditions to use this theoretical dist'n are

The expected counts are all at least 5.

That condition is met here.

# Dataset C output Theoretical Dist'n



# Conclusion

Ho: Location is not associated with flavors of ice cream sold.
Ha: Location is associated with flavors of ice cream sold.

The p-value from the theoretical dist'n is 0.054. This provides moderate evidence in favor of Ha.

The data provide moderate evidence that there is an association between the store and the distribution of flavors of ice-cream.