Name_____

AP Statistics CED 8.1 Daily Video 1

Introducing Statistics – Are My Results Unexpected?

What Will We Learn?

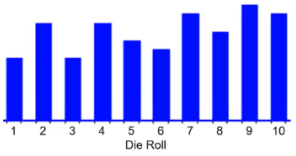
How can we determine if observed counts in categorical data are consistent with expected counts due to random variation?

More Categorical Data

In Unit 6, you learned about inference procedures for categorical data that were classified in terms of _____ and _____. But what is a categorical variable is recorded with _____ _____ categories? In Unit 8, we will investigate inference procedures for the distribution of _____ categorical variable and for the relationship between _____ categorical variables.

Fair Die?

Result	1	2	3	4	5	6	7	8	9	10
Frequency	7	11	7	11	9	8	12	10	13	12



While playing a role-playing game with some friends, you notice that the 10-sided die has been rolling consistently high numbers for a while. Is it your imagination, or could the die be weighted? You decide to roll the die 100 times and record the results. Are these results consistent with what we would expect in random variation?

Expected Counts

In order to answer our question about whether the die is weighted, we must first consider how many of each data value we would _____to see if the die is _____. If the die is fair (unweighted), we would expect to see each of the 10 _____ _____ represented in 100 rolls: _____ = 10 each

Results	1	2	3	4	5	6	7	8	9	10	Total
Observed	7	11	7	11	9	8	12	10	13	12	100
Expected	10	10	10	10	10	10	10	10	10	10	100

How can we determine if the observed counts are a good fit to what we expected based on random variation?

Observed – Expected

If we look at the deviations of the _____ values from the _____ values, we can see where there are large discrepancies. If the die were fair, we would _____ these differences to be close to _____.

Results	1	2	3	4	5	6	7	8	9	10	Total
Observed	7	11	7	11	9	8	12	10	13	12	100
Expected	10	10	10	10	10	10	10	10	10	10	100
Obs – Exp											

But we have a problem: If we try to summarize these differences by taking the sum, this gives us _____!

STATS MEDIC

Name_____

Observed – Expected

Let's try a different approach. What about taking the _____ of the differences to keep the values _____? Though at first this seems like a wise approach, the _____ will be much _____ even if the discrepancies are similar.

Results	1	2	3	4	5	6	7	8	9	10	Total
Observed	7	11	7	11	9	8	12	10	13	12	100
Expected	10	10	10	10	10	10	10	10	10	10	100
Obs – Exp	-3	1	-3	1	-1	-2	2	0	3	2	0
|Obs – Exp|											

Observed – Expected

Another method is to _____. The values will again be _____, but this procedures a similar problem as using the absolute values because with a larger _____, the sum will be much larger.

Results	1	2	3	4	5	6	7	8	9	10	Total
Observed	7	11	7	11	9	8	12	10	13	12	100
Expected	10	10	10	10	10	10	10	10	10	10	100
Obs – Exp	-3	1	-3	1	-1	-2	2	0	3	2	0
$(\text{Obs} - \text{Exp})^2$											

The benefit? This gives _____ differences more weight in their _____ to the sum allowing us to _____ when variation may not be due _____.

Relativity Matters!

Taking into account the _____ of the sample, we can divide each of the _____ differences by their _____.

Results	1	2	3	4	5	6	7	8	9	10	Total
Observed	7	11	7	11	9	8	12	10	13	12	100
Expected	10	10	10	10	10	10	10	10	10	10	100
Obs – Exp	-3	1	-3	1	-1	-2	2	0	3	2	0
$(\text{Obs} - \text{Exp})^2$	9	1	9	1	1	4	4	0	9	4	42
$\frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$											

This results in _____ that better represent the _____ of the differences contributed by the _____ values _____ to what is _____.

The Chi-Square Statistic

This final sum of these _____ is what we call the _____ or _____. Back to the original question…Are these results consistent with what we would _____ in random variation?

Simulating the χ^2 Statistic

http://www.rossmanchance.com/applets/2021/gof/GOF.html

We will be using the online statistical applet *Analyzing One-way Tables* to investigate if the observed counts are a _____ to what we expected based on random variation.

What Should We Take Away?

How can we determine if observed counts in categorical data are consistent with expected counts due to random variation? **We can use the chi-square statistic to measure the distance between the _____ and _____ counts relative to _____ counts.**

STATS MEDIC

Name_____

AP Statistics CED 8.2 Daily Video 1 (Skill 3.C)

Setting up a Chi-Squared goodness of Fit Test

What Will We Learn?

What does the chi-square statistic measure?

How do the degrees of freedom affect the shape of the chi-square distributions?

Predatory Lending

Payday loans and title loans are examples of predatory lending where borrows are left paying interest rates of 100% or more in some cases, making it impossible to pay off the loans. Often, these borrowers have few options due to low credit scores or financial hardship. The "predatory" label of these tactics comes from. Pattern of targeting people who are low-income, elderly, or who have little formal education. Predatory lenders argue that they provide a service and that people from all backgrounds use their services, but is that true?

Predatory Lending

A random sample of 40 predatory lending businesses with Dallas, TX addresses was selected. Their approximate locations are plotted as black points on the map. Do these types of businesses tend to be located primarily in lower income regions, or are they found proportionally in regions of all income levels?

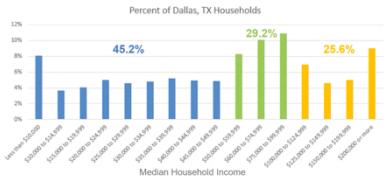
Observed Counts

Household incomes were divided into three categories and the number of predatory lending businesses found in census tracts categorized by the three income brackets was determined.

Median Household Income	Number of Predatory Lending Business
$0 to less than $50,000	20
$50,000 to less than $100,000	17
$100,000 and above	3

Expected Counts

We will use the assumption that the number of businesses found in census tracks corresponding to the three income brackets is _____ to the number of households in the three income brackets.



Observed & Expected Counts

Then we find the _____ between the observed and expected counts. Recall that if we find the sum of these differences, we get _____! (Note: percentages come from the graph on the previous slide.

Median Household Income	Number of Predatory Lending Business	Expected Number of Predatory Lending Businesses	Difference (Obs – Exp)
$0 to less than $50,000	20		
$50,000 to less than $100,000	17		
$100,000 and above	3		

The Chi-Square Statistic

We need something _____, so we _____ the differences to keep all the values _____. Next, we will compare the size of the squared differences _____ _____.

Median Household Income	Number of Predatory Lending Business	Expected Number of Predatory Lending Businesses	Difference (Obs – Exp)	Squared Difference (Obs – Exp)	$\frac{(Obs - Exp)^2}{Exp}$
$0 to less than $50,000	20				
$50,000 to less than $100,000	17				
$100,000 and above	3				

STATS MEDIC

Name_____

The Chi-Square Statistic

Lastly, we will _____ all of the _____ of the observed counts toward the chi-square statistic. If we were thinking in terms of a normal distribution or the *t*-distribution, a test statistic as large as _____ seems rather extreme.

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

$\frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$
0.20389
2.42315
5.11891

Chi-Square Distributions

Similar to *t*-distributions, there are _____ many chi-square distributions whose shapes are determined by the number of _____. However, for inference about the distribution of a single categorical variable, use a chi-square distribution with _____.
NOTE: Degrees of freedom for chi-square distributions are _____ based on _____ size.

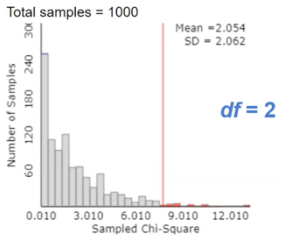
Chi-Square Distributions

To investigate the shape of the distributions of the chi-square statistic with *df*=2, we will simulate drawing a random sample of size 40 from a population where the proportions of predatory lending businesses found in the census tracts from the three income brackets are the same as the proportions of the households in those three income brackets using the online statistical applet *Analyzing One-way Tables* found at: http://www.rossmanchance.com/applets/2021/gof/GOF.html.

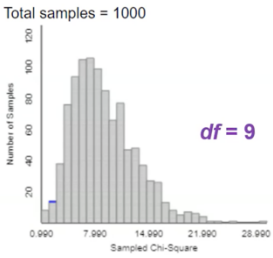
In the simulation after running 1000 simulations only ____/1000 (_____) were ≥ 7.746

Chi-Square Distribution Shape

Here is the simulated distribution of the chi-square statistic for the predatory lending businesses across the _____ when taking random samples of size 40.

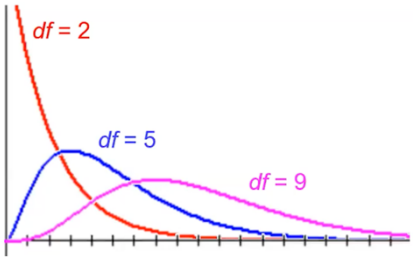


From the previous video, we simulated the distribution of the chi-square statistic when rolling a fair _____ 100 times.



Chi-Square Distribution Shape

We see that as the degrees of freedom _____, the skewness of the chi-square distribution is less pronounced. Additionally, we can see that chi-square distributions have only _____ values.



What Should We Take Away?

What does the chi-square statistic measure?
It measures the distance between _____ and _____ counts _____ to expected counts. Values of the chi-square statistic are always _____.
How do the degrees of freedom affect the shape of the chi-square distributions?
The _____ shape of a chi-square distribution becomes _____ pronounced as the degrees of freedom _____.

STATS MEDIC

Name_____

AP Statistics CED 8.2 Daily Video 2 (Skill 1.F)

Setting Up a Chi-Square Goodness of Fit Test

What Will We Learn?

How do we state the null hypothesis for a chi-square goodness-of-fit test?

How do we state the alternative hypothesis for the chi-square goodness-of-fit test?

Predatory Lending

Payday loans and title loans are examples of predatory lending where borrows are left paying interest rates of 100% or more in some cases, making it impossible to pay off the loans. Often, these borrowers have few options due to low credit scores or financial hardship. The "predatory" label of these tactics comes from. Pattern of targeting people who are low-income, elderly, or who have little formal education. Predatory lenders argue that they provide a service and that people from all backgrounds use their services, but is that true?

Predatory Lending

A random sample of 40 predatory lending businesses with Dallas, TX addresses was selected. Their approximate locations are plotted as black points on the map. Do these types of businesses tend to be located primarily in lower income regions, or are they found proportionally in regions of all income levels?

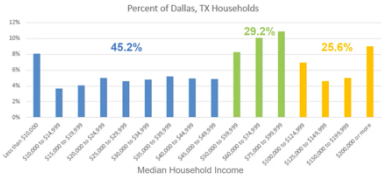
Observed Counts

Household incomes were divided into three categories and the number of predatory lending businesses found in census tracts categorized by the three income brackets was determined.

Median Household Income	Number of Predatory Lending Business
$0 to less than $50,000	20
$50,000 to less than $100,000	17
$100,000 and above	3

Hypothesized Proportions

We will use the assumption that the number of businesses found in census tracks corresponding to the three income brackets is _____ to the number of households in the three income brackets.



Null Hypothesis

When we write the null hypothesis for a chi-square goodness-of-fit test, we can write it in _____ or _____.

In words: H_0: The distribution of _____ in Dallas, TX is the same as the _____ of households in the specified income brackets.

In symbols: H_0: p_1 = _____, p_2 = _____, p_3 = _____

where p_1, p_2 and p_3 represent the _____ of predatory lending businesses found in regions of Dallas, TX where household incomes are _____, _____ _____ and _____, respectively.

Alternative Hypothesis

Recall that we need a form of _____ in the alternative hypothesis. We have _____ proportions to consider. Since these proportions come from categories that cover _____ possible outcomes, their sum will be _____. If _____ from its hypothesized proportion, that means at least _____ will also be different since all the proportions must sum to _____.

p_1 = _____, p_2 = _____, p_3 = _____. Sum of proportions = _____

Name_____

Alternative Hypothesis

You should _____ state the alternative hypothesis in a way that suggests _____ the proportions in the null hypothesis are different. ✗ H_a: $p_1 \neq .452$, $p_2 \neq .292$, $p_3 \neq .256$

We are testing for _____, so given convincing evidence that at _____ proportion is different from what is stated in the null hypothesis is _____ to show the data do not _____. Likewise, you should _____ state the alternative hypothesis with directional _____ like < or > that we used for significance testing in previous units.

Remember: State the _____ in _____!!

Summing Up Hypotheses

For hypotheses about a distribution of proportions for a _____ categorical variable:

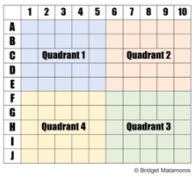
- The null hypothesis is a statement of _____ where the proportions are all equal to specified values.
- The alternative hypothesis states that _____ on proportion is not as specified in the null hypothesis.
- Never refer to _____ proportions (such as _____) in the hypotheses!
- Remember to define any _____ or _____ you use.

Battleship Quadrants

Are certain quadrants on the Battleship game field preferred by players as strategic locations for their ships? A random sample of 100 people who enjoy the game Battleship were surveyed. They were asked to place their five ships on a Battleship game field as if they were going to begin playing against a worthy opponent. The quadrant of the game field with the greatest number of spaced occupied by ships was recorded. Ties were broken by identifying the greatest number of ships found in or near a quadrant. Quadrant 1 was defined as the northwest corner (containing space A1) and subsequent quadrants were identified in clockwise order.

Battleship Quadrants

Quadrant	1	2	3	4
Observed	16	22	33	29



The observed counts for each quadrant are found in the table below. Is there convincing evidence that players have a preference for certain quadrants of the Battleship game field? State the hypotheses for a chi-square goodness-of-fit test.

Null Hypothesis

We are not given a specified distribution to assume for the proportion of players that prefer each of the four quadrants. Instead, if players have _____ for where to place their ships, we would _____ to see all the quadrants _____ represented.

H_0: The _____ of Battleship quadrants preferred by players is the same across _____ _____ quadrants, _____. OR H_0: $p_1 = p_2 = p_3 = p_4 = 0.25$ where p_1, p_2, p_3, and p_4 are the _____ of players that place the most ships in Battleship quadrant 1, 2, 3, and 4.

Alternative Hypothesis (Remember to write this in WORDS!)

H_a: The distribution of Battleship quadrants preferred by players is not the same across _____ quadrants OR H_a: _____ of the proportions is not as specified in the _____ hypothesis.

What Should We Take Away?

How do we state the null hypothesis for a chi-square goodness-of-fit test? **The null hypothesis states that the proportions for the categories in a _____ categorical variable are _____ to specified values.**

How do we state the alternative hypothesis for the chi-square goodness-of-fit test? **In _____, state that at _____ of the proportions is not as specified in the _____ hypothesis.**

Name_____

AP Statistics CED 8.2 Daily Video 3 (Skill 4.C)

What Will We Learn?

How do we identify an appropriate significance test procedure for a distribution of proportions for one categorical variable?

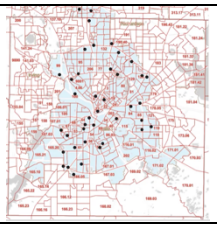
How do we check the conditions for performing a significance test for a distribution of proportions for one categorical variable?

Predatory Lending

Payday loans and title loans are examples of predatory lending where borrows are left paying interest rates of 100% or more in some cases, making it impossible to pay off the loans. Often, these borrowers have few options due to low credit scores or financial hardship. The "predatory" label of these tactics comes from. Pattern of targeting people who are low-income, elderly, or who have little formal education. Predatory lenders argue that they provide a service and that people from all backgrounds use their services, but is that true?

Predatory Lending

A random sample of 40 predatory lending businesses with Dallas, TX addresses was selected. Their approximate locations are plotted as black points on the map. Do these types of businesses tend to be located primarily in lower income regions, or are they found proportionally in regions of all income levels?



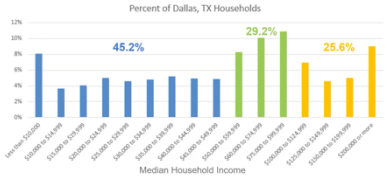
Observed Counts

Household incomes were divided into three categories and the number of predatory lending businesses found in census tracts categorized by the three income brackets was determined.

Median Household Income	Number of Predatory Lending Business
$0 to less than $50,000	20
$50,000 to less than $100,000	17
$100,000 and above	3

Hypothesized Proportions

We will use the assumption that the number of businesses found in census tracks corresponding to the three income brackets is _____ to the number of households in the three income brackets.



Hypotheses

H_0: p_1 = _____, p_2 = _____, p_3 = _____

where p_1, p_2 and p_3 represent the _____ of predatory lending businesses found in regions of Dallas, TX where household incomes are _____, _____ _____ and _____, respectively.

H_a: _____ of the proportions is _____ as specified in the _____ hypothesis.

Identifying the Procedure

What type of data were collected? Business locations were recorded and _____ across three _____ of _____ variable: median household income bracket.

How many groups? We have _____ sample of _____ predatory lending businesses.

What are we asked to do? Determine if there is _____ that predatory lending businesses are _____ located in regions of _____ income level _____.

Chi-Square Goodness-of-Fit Test

STATS MEDIC

Name_____

Checking the Conditions

To check for independence

1. The data should come from a _____ sample OR a _____ experiment.
2. When sampling _____, the sample should be less than or equal to _____ of the respective population.

The chi-square goodness-of-fit test becomes _____ with more observations, so _____ should be greater than _____.

Checking the Conditions (Be sure to ✓ your conditions!)

1. A _____ predatory businesses was selected.
2. We will assume that _____ lending businesses is less then or equal to _____ of _____ lending businesses in Dallas, TX.

To verify that the expected counts are all _____, we need the proportions from the _____ hypothesis: p_1 = _____, p_2 = _____, p_3 = _____

Expected Counts

We find the expected counts by multiplying the proportions of each category from the null hypothesis with the sample size. (Complete table!)

3. All expected _____.

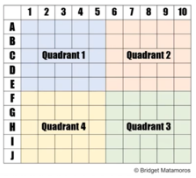
Median Household Income	Number of Predatory Lending Business	Expected Number of Predatory Lending Businesses
$0 to less than $50,000	20	
$50,000 to less than $100,000	17	
$100,000 and above	3	

Battleship Quadrants

Are certain quadrants on the Battleship game field preferred by players as strategic locations for their ships? A random sample of 100 people who enjoy the game Battleship were surveyed. They were asked to place their five ships on a Battleship game field as if they were going to begin playing against a worthy opponent. The quadrant of the game field with the greatest number of spaced occupied by ships was recorded. Ties were broken by identifying the greatest number of ships found in or near a quadrant. Quadrant 1 was defined as the northwest corner (containing space A1) and subsequent quadrants were identified in clockwise order.

Battleship Quadrants

Quadrant	1	2	3	4
Observed	16	22	33	29



The observed counts for each quadrant are found in the table below. Is there convincing evidence that players have a preference for certain quadrants of the Battleship game field? Identify the procedure and verify that the conditions for inference have been met.

Hypotheses

H_0: $p_1 = p_2 = p_3 = p_4$ = _____; where p_1 , p_2 , p_3 , and p_4 are the _____ of players that place the most ships in Battleship quadrant 1, 2, 3, and 4.

H_a: _____ of the proportions is _____ as specified in the _____ hypothesis.

Identifying the Procedure

What type of data were collected? People were recorded as _____ counts across _____ _____ of _____: which of the four quadrants had the most ships.

How many groups? We have _____ sample of _____ people who enjoy playing Battleship.

What are we asked to do? Determine if there is _____ that people prefer to place their ships in certain quadrants of the Battleship game field.

Chi-Square Goodness-of-Fit Test

STATS MEDIC

Name_____

Checking the Conditions (Be sure to ✓ your conditions!)

1. A _____ of _____ who enjoy playing the game Battleship was selected.
2. There are at least _____ = _____ people who enjoy playing the game Battleship.

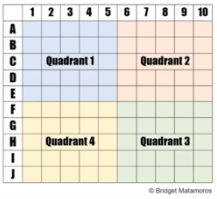
Battleship Quadrants (Be sure to ✓ your conditions!)

To verify that the expected counts are all _____, we need the proportions stated in the _____ hypothesis.

H_0: $p_1 = p_2 = p_3 = p_4 =$_____

3. All _____ are _____.

Quadrant	1	2	3	4
Observed	16	22	33	29
Expected	**25**	**25**	**25**	**25**



What Should We Take Away?

How do we identify an appropriate significance test procedure for a distribution of proportions for one categorical variable?

Use a _____

How do we check the conditions for performing a significance test for a distribution of proportions for one categorical variable?

1. The data should come from a _____ or a _____ experiment.

2. When samples _____, the sample should be _____ of the respective population.

3. All _____ should be _____.

Name_____

AP Statistics CED 8.3 Daily Video 1 (Skill 3.E)

Carrying Out a Chi-Square Goodness-of-Fit Test

What Will We Learn?

How do we calculate a test statistic for a chi-square goodness-of-fit test?

How do we calculate a *p*-value for a chi-square goodness-of-fit test?

Predatory Lending Recap

Payday loans and title loans are examples of predatory lending where borrows are left paying interest rates of 100% or more in some cases, making it impossible to pay off the loans. Often, these borrowers have few options due to low credit scores or financial hardship. The "predatory" label of these tactics comes from. Pattern of targeting people who are low-income, elderly, or who have little formal education. Predatory lenders argue that they provide a service and that people from all backgrounds use their services, but is that true?

Predatory Lending Recap

A random sample of 40 predatory lending businesses with Dallas, TX addresses was selected. Their approximate locations are plotted as black points on the map. Do these types of businesses tend to be located primarily in lower income regions, or are they found proportionally in regions of all income levels?

Observed Counts

Household incomes were divided into three categories and the number of predatory lending businesses found in census tracts categorized by the three income brackets was determined.

Median Household Income	Number of Predatory Lending Business
$0 to less than $50,000	20
$50,000 to less than $100,000	17
$100,000 and above	3

Hypotheses and Conditions

Recall from the previous video that the hypotheses for this chi-square goodness-of-fit test are:

H_0: p_1 = _____, p_2 = _____, p_3 = _____; where p_1, p_2 and p_3 represent the _____ of predatory lending businesses found in regions of Dallas, TX where household incomes are _____, _____ and _____, respectively.

H_a: _____ is not as specified in the _____.

Additionally, the conditions for inference have _____.

Our next step is to proceed with the mechanics of the test:

calculating the _____ and the _____.

Calculating the Test Statistic (Complete the table below as you watch the video.)

In the previous video, we found the expected counts by _____ the proportions of each category from the _____ hypothesis by the _____. Next we need to calculate the _____ between the _____ and _____ counts. To keep the values _____ as well as give _____ to extreme deviations, we will _____ the differences. Taking into account the _____ of the sample, we will find the _____ of the _____ differences and the _____ _____.

Median Household Income	Number of Predatory Lending Business	Expected Number of Predatory Lending Businesses	Difference (Obs – Exp)	Squared Difference (Obs – Exp)	$\frac{(Obs - Exp)^2}{Exp}$
$0 to less than $50,000	20				
$50,000 to less than $100,000	17				
$100,000 and above	3				

Name_____

Calculating the Test Statistic

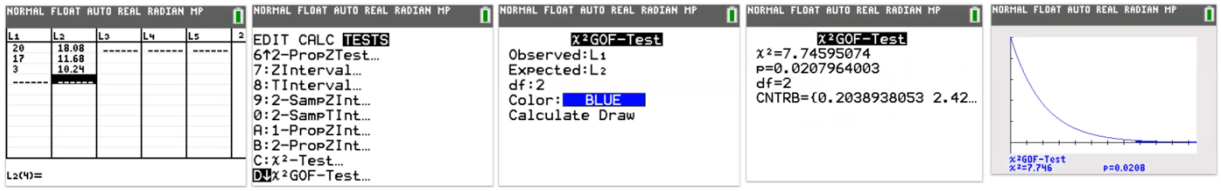
Lastly, we will find the _____ of these ratios.

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

$\frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$
0.20389
2.42315
5.11891

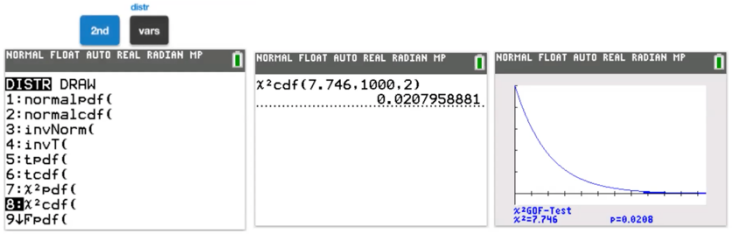
Calculating the *p*-value

We can use technology to calculate the *p*-value. **Remember: *df* = number of categories – 1**



Calculating the *p*-value

You can use the χ^2cdf function on the calculator

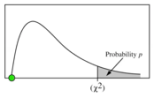


χ^2cdf (lowerbound, upperbound, df)

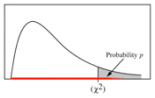
Alternatively, you could use Table C.
(If this is your choice, follow the video.)

Why the Upper Tail?

A chi-square statistic of _____ means the observed and expected counts are _____ (A perfect fit!).



As we move further _____ from zero, this implies the _____ between the observed and expected counts are _____ which means the _____ is _____.



Battleship Quadrants

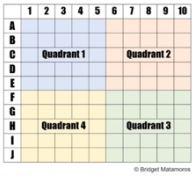
Are certain quadrants on the Battleship game field preferred by players as strategic locations for their ships? A random sample of 100 people who enjoy the game Battleship were surveyed. They were asked to place their five ships on a Battleship game field as if they were going to begin playing against a worthy opponent. The quadrant of the game field with the greatest number of spaced occupied by ships was recorded. Ties were broken by identifying the greatest number of ships found in or near a quadrant. Quadrant 1 was defined as the northwest corner (containing space A1) and subsequent quadrants were identified in clockwise order.

Battleship Quadrants

The observed counts for each quadrant are found in the table below. Pause the video and use the space below to calculate the test-statistic and *p*-value.

Quadrant	1	2	3	4
Observed	16	22	33	29
Expected				

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



STATS MEDIC

Name_____

Multiple-Choice Example

Which of the following gives the correct test statistic and *p*-value for a chi-square goodness-of-fit test to determine if there is convincing evidence that the players have a preference for certain quadrants of the Battleship game field? (Eliminate choices as you watch the video!)

(A) $\chi^2 = \frac{(25-16)}{16} + \frac{(25-22)}{22} + \frac{(25-33)}{33} + \frac{(25-29)}{29}$; *p*-value = 0.957

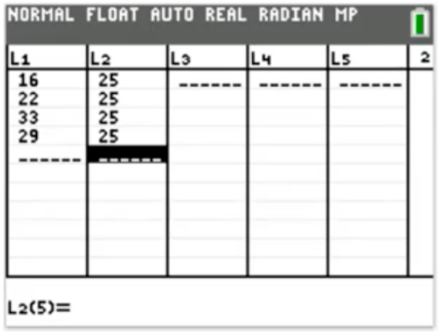
(B) $\chi^2 = \frac{(16-25)}{25} + \frac{(22-25)}{25} + \frac{(33-25)}{25} + \frac{(29-25)}{25}$; *p*-value = 1.0

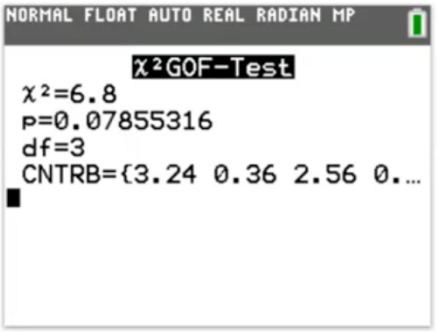
(C) $\chi^2 = \frac{(16-25)^2}{25} + \frac{(22-25)^2}{25} + \frac{(33-25)^2}{25} + \frac{(29-25)^2}{25}$; *p*-value = 0.147

(D) $\chi^2 = \frac{(16-25)^2}{25} + \frac{(22-25)^2}{25} + \frac{(33-25)^2}{25} + \frac{(29-25)^2}{25}$; *p*-value = 0.079

(E) $\chi^2 = \frac{(16-25)^2}{100} + \frac{(22-25)^2}{100} + \frac{(33-25)^2}{100} + \frac{(29-25)^2}{100}$; *p*-value = 0.637

Multiple-Choice Example





What Should We Take Away?

How do we calculate a test statistic for a chi-square goodness-of-fit test?

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

How do we calculate a *p*-value for a chi-square goodness-of-fit test?

Calculate the _____ probability from a chi-square distribution with _____ using technology.

STATS MEDIC

Name_____

AP Statistics CED 8.3 Daily Video 2

Carrying Out a Chi-Square Goodness-of-Fit Test

What Will We Learn?

How do we interpret the *p*-value for a chi-square goodness-of-fit test?
How so we state a conclusion for a chi-square goodness-of-fit test?

Predatory Lending Recap

Payday loans and title loans are examples of predatory lending where borrows are left paying interest rates of 100% or more in some cases, making it impossible to pay off the loans. Often, these borrowers have few options due to low credit scores or financial hardship. The "predatory" label of these tactics comes from. Pattern of targeting people who are low-income, elderly, or who have little formal education. Predatory lenders argue that they provide a service and that people from all backgrounds use their services, but is that true?

Predatory Lending Recap

A random sample of 40 predatory lending businesses with Dallas, TX addresses was selected. Their approximate locations are plotted as black points on the map. Do these types of businesses tend to be located primarily in lower income regions, or are they found proportionally in regions of all income levels?

Observed Counts

Household incomes were divided into three categories and the number of predatory lending businesses found in census tracts categorized by the three income brackets was determined.

Median Household Income	Number of Predatory Lending Business
$0 to less than $50,000	20
$50,000 to less than $100,000	17
$100,000 and above	3

Predatory Lending Recap

Hypotheses: **H_0: $p_1 = .452$, $p_2 = .292$, $p_3 = .256$; where p_1, p_2 and p_3 represent the proportions of predatory lending businesses found in regions of Dallas, TX where household incomes are $0 to less than $50,000, $50,000 to less than $100,000 and $100,000 and above, respectively.**
H_a: At least one of the proportions is not as specified in the null hypothesis.
Procedures & Conditions: **The conditions for a chi-square goodness-of-fit test have been verified.**
Mechanics: **The test statistics is $\chi^2 = 7.746$ and the *p*-value = 0.0208**

Interpreting a *p*-value

Recall from the previous video that the *p*-value is calculated by finding the _____ tail probability of a chi-square distribution. What exactly does this describe? From Units 6 and 7, we learned that the _____ of obtaining a result as _____ the one in the study, or _____ by chance alone, _____ the null hypothesis is _____.

Interpreting a *p*-value

There is a _____ of getting a chi-square statistic of _____ just by the chance involved in the _____ selection of the businesses, _____ that the distribution of predatory lending businesses in Dallas, TX is the _____ as the proportions of households in the specific _____.

Median Household Income	Number of Predatory Lending Businesses	Expected Number of Predatory Lending Businesses	Obs – Exp	(Obs – Exp)²	$\frac{(Obs - Exp)^2}{Exp}$
$0 to less than $50,000	20	18.08	1.92	3.6864	**0.20389**
$50,000 to less than $100,000	17	11.68	5.32	28.3024	**2.42315**
$100,000 and above	3	10.24	–7.24	52.4176	**5.11891**
					$\chi^2 = 7.746$

Name_____

Stating a Conclusion: A General Guide

When we state our conclusion, it has _____ main parts:

1. How does the *p*-value _____ to our level of significance, α, and what _____ must be make about H_0?
2. What does this mean about H_a in _____?

For small *p*-values ⟶ test statistics is _____ to occur by random _____.

Since the *p*-value of _____ ≤ α = _____, we _____ H_0.

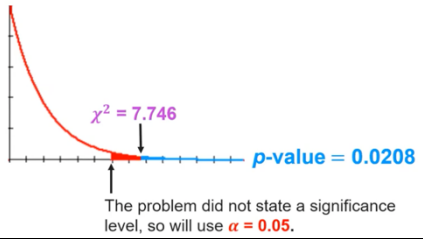
There is _____ statistical evidence that [state H_a in _____]

For large *p*-values ⟶ test statistics is _____ to occur by random _____.

Since the *p*-value of _____ > α = _____, we _____ H_0.

There is _____ statistical evidence that [state H_a in _____]

Stating a Conclusion



Since the *p*-value of _____ is _____ α = _____, we _____ the H_0.

There _____ convincing _____ that the distribution of predatory lending businesses in Dallas, TX is _____ as the proportions of households in the specified income brackets.

Contributions

Let's return to the original question: Is there convincing statistical evidence that these types of businesses tend to be located primarily in lower income regions? We have determined that the proportions of predatory lending businesses found in regions of different income brackets are not the same as the proportions of households in the specified income brackets.

⟶ **Are the located primarily in lower income regions?**

Contributions

By looking at the _____ of each _____ toward the chi-square statistic, we can determine where the _____ lies in what we observed and what we expected. We can see that the _____ bracket had the largest contribution. From these data, we can infer that predatory lending businesses do tend to be located in _____ regions because they were represented _____ _____ in the highest income regions.

Median Household Income	Number of Predatory Lending Businesses	Expected Number of Predatory Lending Businesses	Obs – Exp	(Obs – Exp)²	$\frac{(Obs - Exp)^2}{Exp}$
$0 to less than $50,000	20	18.08	1.92	3.6864	**0.20389**
$50,000 to less than $100,000	17	11.68	5.32	28.3024	**2.42315**
$100,000 and above	3	10.24	–7.24	52.4176	**5.11891**
					$\chi^2 = 7.746$

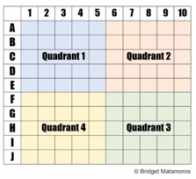
Multiple Choice Example

Are certain quadrants on the Battleship game field preferred by players as strategic locations for their ships? A random sample of 100 people who enjoy the game Battleship were surveyed. They were asked to place their five ships on a Battleship game field as if they were going to begin playing against a worthy opponent. The quadrant of the game field with the greatest number of spaced occupied by ships was recorded. Ties were broken by identifying the greatest number of ships found in or near a quadrant. Quadrant 1 was defined as the northwest corner (containing space A1) and subsequent quadrants were identified in clockwise order.

Name_____

Battleship Quadrants

Quadrant	1	2	3	4
Observed	16	22	33	29



The observed counts for each quadrant are found in the table below. A chi-square goodness-of-fit test to determine whether the proportions of players preferring each of the four quadrants are the same resulted in a *p*-value of 0.079.

Multipe-Choice Example. (Eliminate answers as you watch the video!)

Which of the following is a correct interpretation of the *p*-value? ← **PAY ATTENTION TO WHAT THE QUESTION IS ASKING!**

(A) Since the *p*-value of 0.079 is greater than $\alpha = .05$, we should reject the null hypothesis that the proportions are the same. There is convincing evidence that players prefer certain quadrants.

(B) Three of the observed counts are not greater than 30, so the conditions have not been met. An interpretation of the *p*-value would not be valid.

(C) Only 7.9% of those surveyed preferred quadrant 1. This is convincing evidence that the proportions of players preferring different quadrants is not the same.

(D) There is a 0.079 probability that we would get results this extreme or more extreme, by chance alone, if the proportions of players preferring different quadrants is the same.

(E) There is a 0.079 probability that we would get results this extreme or more extreme, by chance alone, assuming that players have preferences for different quadrants.

What Should We Take Away?

How do we interpret the *p*-value for a chi-square goodness-of-fit test?

Assuming the proportions of a _____ categorical variable as stated in the _____ hypothesis are _____, there is a <*p*-value> _____ of getting a chi-square statistic as _____ as the one in the study _____, by chance alone in _____ sampling (or _____ assignment).

How so we state a conclusion for a chi-square goodness-of-fit test?

Since the *p*-value of _____ $\leq \alpha =$ _____, we _____ H_0.
There is _____ statistical evidence that [state H_a in _____]

Since the *p*-value of _____ $> \alpha =$ _____, we _____ H_0.
There is _____ statistical evidence that [state H_a in _____]

STATS MEDIC

Name_____

AP Statistics CED 8.3 Daily Video 3 (Skill 4.E)

Carrying Out a Chi-Square Goodness-of-Fit Test

What Will We Learn?

How do we perform a complete significance test for a distribution of proportions for one categorical variable?

2008 International Exam (The first section was omitted from the video, provided here for clarity!)

"The department of parks and recreation of a certain city conducts summer programs for residents of its six districts. The summer programs include operating and maintaining community swimming pools in each of the districts as well as offering sports and recreational programs for school-age children, young adults, and older adults. The table above shows the proportion of households by district out of all the households that participated in the summer programs, based on annual data that were collected from simple random samples each summer over a 10-year period, ending in the year 2000. The proportions are being used by the city for planning purposes and for more efficiently targeting the introduction of future programs."

District	A	B	C	D	E	F
Proportion of Households	0.32	0.12	0.10	0.27	0.05	0.14

City leaders want to test if the proportions that are being used by the city are still valid. Data collected by a statistician from a simple random sample this past summer indicated that the following **number** of households participated in each district.

District	A	B	C	D	E	F
Number of Households	100	35	40	22	12	31

(a) The statistician claims that the data for this past summer provide evidence that the proportions that are being used by the city are no longer valid. Give statistical evidence to justify the claim.

(b) Which one of the six districts had the greatest change in participation since the year 2000? Use the information from part (a) to explain your choice.

Planning the Response (Highlight the key information above as you watch the video.)

First, this is a _____ test because we are asked to give _____ evidence for a claim. Let's identify the procedure, we have _____ sample and the variable of interest is a _____ variable with _____ categories. This tells us we will need to perform a _____.

Part (a) – Significance test procedures generally require 4 parts:

- _____
- _____
- _____
- _____

Part (b) – We need the contributions!

Hypotheses & Parameters

Here are the current proportions being used by City Leaders that are being challenged by the researchers.

District	A	B	C	D	E	F
Proportion of Households	0.32	0.12	0.10	0.27	0.05	0.14

H_0: p_A = _____, p_B = _____, p_C = _____, p_D = _____, p_E = _____ p_F = _____ where p_A through p_F are the _____of households participating in the summer program for _____ district.

H_a: _____ of the proportions is different from those specified in the _____ hypothesis.

Procedures & Conditions (Be sure to ✓ your conditions!)

Identification of procedure: **We will use a chi-square _____.**

Conditions:

- **The study uses a _____ sample of households.**
- **We must assume that _____ households are less than or equal to ____ of all households in the city.**

STATS MEDIC

Name_____

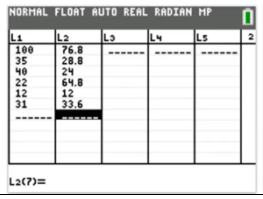
If we multiply each proportion by the _____ of ____, you will get the expected values. You do need to state these on your AP Exam!

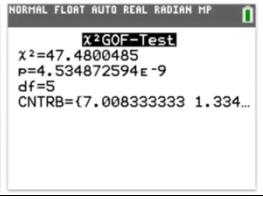
3. All expected counts are _____.

District	A	B	C	D	E	F
Proportion of Households	0.32	0.12	0.10	0.27	0.05	0.14
Expected Number of Households						

Mechanics (Use Technology!) (Note: Pay close attention to the ***df***: number of categories – 1 or 6 – 1 = 5)

Enter Lists | Select χ^2GOF-Test | Assign Data | Calculate | Results





You must record all three of these values on the AP Exam!

$\chi^2 = 47.48$

p-value = $4.53 \times 10^{-9} \approx 0$

$df = 5$

Conclusion

Since the p-value of approximately _____ is _____ α = _____, we _____ the _____. There ___ convincing _____ that the proportions of households participating in the _____ for each district are _____ the same as those being use by the city.

Part (b) Which one of the six districts had the greatest change in participation since the year 2000? Use the information from part (a) to explain your choice.

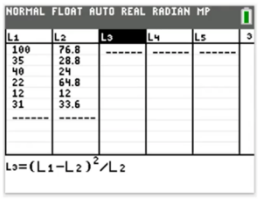
We can compute this by hand using the table and then find the category that made the largest contribution. District _____

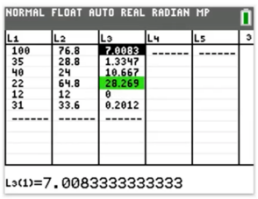
District	A	B	C	D	E	F
Proportion of Households	0.32	0.12	0.10	0.27	0.05	0.14
Expected Number of Households	76.8	28.8	24	64.8	12	33.6
Number of Households	100	35	40	22	12	31
$\frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$	7.0083	1.3347	10.6667	28.2691	0	0.2012

Test Tip: List Operations

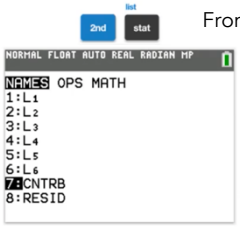
You can calculate these contributions using technology.

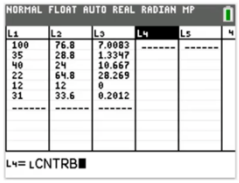
Using the formula in Lists

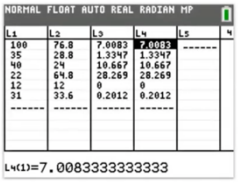




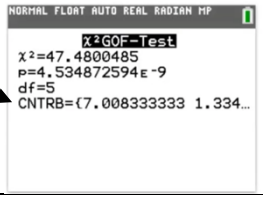
From the stored List @ 2nd List







From the χ^2GOF-Test Results



District _____ had the greatest _____ in Participation because it has the _____ contribution (CNTRB) toward the value of the chi-square statistic (_____).

Scoring Guidelines

Section 1: State a correct pair of _____ and check the _____.

Section 2: Show the correct mechanics including the value of the _____, ____, and _____

Section 3: State a correct _____ in _____ using the _____ of the test.

Section 4: Identify and show supporting _____ for the selection of the correct district (D).

STATS MEDIC

Name_____

What Should We Take Away?

How do we perform a complete significance test for a distribution of proportions for one categorical variable?

Be sure to:

- ☐ **Define any _____ and _____ used.**
- ☐ **State the _____ - Always specify value for the _____ in each category.**
- ☐ **Identify the _____ you are using.**
- ☐ **Verify that the _____ for the procedure have been _____.**
- ☐ **Calculate the _____, _____, and _____.**
- ☐ **Interpret the _____ in _____>**

Name_____

AP Statistics CED 8.4 Daily Video 1

Expected Counts in Two-Way Tables

What Will We Learn?

How do we calculate the expected counts involving two-way tables of categorical data?

A Look Forward

In Topics 8.4 – 8.6 we will learn _____ new types of chi-square significance tests.

For both tests, we will be analyzing data in a _____.

A two-way table _____ data for two _____ variables.

		Variable 1		
		Category 1	Category 2	Category 3
Variable 2	Category A			
	Category B			
	Category C			

© Luke Wilcox

Where Do You Go to School?

A random sample of parents with school-aged children was taken during 2019. A separate random sample of parents with school-aged children was taken in 2020. Parents were asked what type of school their children attended. Here are the results.

	2019 Sample	2020 Sample
Public	266	163
Private/Parochial/Charter	16	21
Home	38	30

The variable of _____ is represented in the columns and the variable of _____ is represented in the rows.

Expected Counts

Expected counts in a two-way table are calculated _____ there is no _____ between the two categorical variables.

There are three steps to totaling a two-way table:

- Find row totals
- Find column totals
- Find the table total

	2019 Sample	2020 Sample	Total
Public	266	163	
Private/Parochial/Charter	16	21	
Home	38	30	
Total			

* Overall proportion that attended public school = _____ = _____ (So, assuming no relationship!)
* 80.34% of the 2019 sample should attend public school, or (0.8034)(320) = _____
* expected count (public 2019) = (_____)(_____) = _____

General formula: $\textit{expected count} = \frac{(\quad)(\quad)}{(\quad)}$

Expected Counts (Use the General Formula from above!)

Expected count (Private/Parochial/Charter, 2019) = _____ = _____

Expected count (Home, 2019) = _____

Expected count (Public, 2020) = _____

Expected count (Private/Parochial/Charter, 2020) = _____

Expected count (Home, 2020) = _____

(Note: Make sure you include a key to indicate the number in ( ) are expected counts.)

	2019 Sample	2020 Sample	Total
Public	266 (257.1)	163	429
Private/Parochial/Charter	16	21	37
Home	38	30	68
Total	320	214	534

Name_____

Are you Working?

A random sample of 2000 adults revealed the following data about education level and employment status.

_____ is the categorical variable represented by rows.

_____ is the categorical variable represented by columns.

Calculate the expected counts:

- Start by finding the row totals, column totals and the table total. (Add lines!)
- Use the formula to calculate expected counts for each cell. (Add these values to the table.)
- When possible, use subtraction to find the remaining expected counts.
- Don't forget the Key!

	No High School Diploma	High School Diploma, No College	High School Diploma, Some College
Employed	206	548	1186
Unemployed	14	22	24

What Should We Take Away?

How do we calculate the expected counts involving two-way tables of categorical data?

$$expected\ count = \frac{(row\ total)(column\ total)}{table\ total}$$

Once we have calculated enough of these expected counts, use the row totals and column totals and subtraction to find the others.

STATS MEDIC

Name_____

AP Statistics CED 8.5 Daily Video 1 (Skill 1.F)

Setting up a Chi-Square Test for Homogeneity or Independence

What Will We Learn?

How do we identify the appropriate significance test procedure for data in a two-way table?
How do we state a null and alternative hypothesis for a chi-square test for homogeneity?
How do we state a null and alternative hypothesis for a chi-square test for independence?

Identifying the Procedure

There are _____ types of significance tests for _____ data in a two-way table.

- **Goal:** to _____ distributions of a categorical variable for _____ populations
- **Data collection:** _____ random samples from _____ populations (or from _____ groups in a randomized experiment)
  → *chi-square test for* _____
- **Goal:** to determine whether two categorical variables are _____
- **Data collection:** _____ random sample from _____ population
  → *chi-square test for* _____
- The appropriate test for data in a two-way table depends on _____ _____.

Where Do You Go To School?

A random sample of parents with school-aged children was taken during 2019. A separate random sample of parents with school-aged children was taken in 2020. Parents were asked what type of school their children attended. Here are the results.

	2019 Sample	2020 Sample
Public	266	163
Private/Parochial/Charter	16	21
Home	38	30

If we are interested in comparing the distribution of type of school for the two year, which significance test is appropriate?

Goal: The goal is to compare the distribution of _____ between the two years
Data collection: _____ independent random samples from _____ different populations
Therefore, the is a chi-square test for _____.

Null Hypothesis

In a statistical test, the _____ hypothesis is often a claim of _____ or _____.

H_0: There is _____ in the distribution of _____ attended by school-aged children from 2019 to 2020. *(Notice that there is no parameter to define.)*

Alternative Hypothesis

In a statistical test, the _____ hypothesis is the _____ that we hope to support with _____ from the data collected.

H_0: There is _____ in the distribution of _____ attended by school-aged children from 2019 to 2020.

H_a: There is _____ in the distribution of _____ attended by school-aged children from 2019 to 2020.

Because the alternative hypothesis includes several proportions that could differ in either direction, this test is said to be _____.

Name_____

Are you Working?

A random sample of 2000 adults revealed the following data about education level and employment status.

	No High School Diploma	High School Diploma, No College	High School Diploma, Some College
Employed	206	548	1186
Unemployed	14	22	24

If we are interested in deciding if education level and employment status are associated, what is the appropriate significance test?

Goal: To see if there is an _____ between two categorical variables.

Data collection: data collected from _____ random sample from _____ population. Also, note the work _____. Therefore, this is a chi-square test for _____.

Null Hypothesis. (Remember, no difference or no change!)

H_0: There is _____ between education level and employment status for _____. OR

H_0: Education level and employment status are _____ for all adults.

Alternative Hypothesis. (Claim we hope to support with evidence from data collected.)

H_0: There is _____ between education level and employment status for _____.

H_a: There is _____ between education level and employment status for _____. OR

H_0: Education level and employment status are _____ for all adults.

H_a: Education level and employment status are _____ for all adults.

What Should We Take Away?

How do we identify the appropriate significance test procedure for data in a two-way table?

_____ ***categorical variable,*** _____ ***populations → chi-square test for*** _____

_____ ***categorical variables,*** _____ ***population → chi-square test for*** _____

How do we state a null and alternative hypothesis for a chi-square test for homogeneity?

H_0: There is _____ in the distribution of [_____] across populations or treatments.

H_a: There is _____ in the distribution of [_____] across populations or treatments.

How do we state a null and alternative hypothesis for a chi-square test for independence?

H_0: There is _____ between [categorical variable] and [categorical variable] in a given population.

H_a: There is _____ between [categorical variable] and [categorical variable] in a given population.

OR

H_0: [categorical variable] and [categorical variable] are _____ in a given population.

H_a: [categorical variable] and [categorical variable] are _____ in a given population.

Name_____

AP Statistics CED 8.5 Daily Video 2 (Skill 4.C)

Setting Up a Chi-Square Test for Homogeneity or Independence

What Will We Learn?

How do we verify the conditions for performing a chi-square test for homogeneity or independence?

Where Do You Go to School?

A random sample of parents with school-aged children was taken during 2019. A separate random sample of parents with school-aged children was taken in 2020. Parents were asked what type of school their children attended. Here are the results.

	2019 Sample	2020 Sample
Public	266	163
Private/Parochial/Charter	16	21
Home	38	30

Researchers would like to perform a test to determine if the distribution of school type for all school-aged children differed between the two years. Check the conditions for inference.

Where Do You Go to School?

In a previous video, we stated the hypotheses:

H_0: There is _____ in the distribution of school types attended by school-aged children from 2019 to 2020.

H_a: There is _____in the distribution of school types attended by school-aged children from 2019 to 2020.

We also identified the procedure as a *chi-square test for* _____.

Checking the Conditions

Remember that for _____ procedures in AP Statistics you _____ verify that the _____ for using that procedure are _____.

In general, you should check for:

- independence in the methods used to _____ the data, and
- that the appropriate _____ has the correct shape.

Checking the Conditions

Here are the conditions for a chi-square test for homogeneity or independence.

To check for _____:

1. Data should be collected using a _____ random sample or randomized experiment (_____) or a _____ random sample (_____)
2. When sampling _____ replacement, the sample size is _____ to 10% of the population size.

To check that the shape of the _____ distribution is approximately a _____ distribution:

3. All _____ counts should be _____.

Name_____

Where Do You Go to School?

A random sample of parents with school-aged children was taken during 2019. A separate random sample of parents with school-aged children was taken in 2020. Parents were asked what type of school their children attended. Here are the results.

	2019 Sample	2020 Sample
Public	266	163
Private/Parochial/Charter	16	21
Home	38	30

Researchers would like to perform a test to determine if the distribution of school type for all school-aged children differed between the two years.

Check the conditions for inference. (Be sure to ✓ your conditions!)

1. There are _____ random samples (_____ random samples).
2. It is _____ to believe _____ parents is _____ to _____ or all parents in 2019 and _____ parents is _____ to _____ or all parents in 2020.
3. All expected counts are _____.

(In a previous video we calculated the expected counts.)

	2019 Sample	2020 Sample
Public	266 (257.1)	163 (171.9)
Private/Parochial/Charter	16 (22.2)	21 (14.8)
Home	38 (40.7)	30 (27.3)
Total	320	214

(expected counts)

Are you Working?

A random sample of 2000 adults revealed the following data about education level and employment status. Researchers want to determine if there is an association between education level and employment status. Check the conditions for inference.

	No High School Diploma	High School Diploma, No College	High School Diploma, Some College
Employed	206	548	1186
Unemployed	14	22	24

Are you Working?

In a previous video, we stated the hypotheses:

H_0: There is no association between education level and employment status for all adults.

H_a: There is an association between education level and employment status for all adults.

We also identified the procedure as a *chi-square test for* _____.

Are you Working? (Check the conditions using information above and ✓ your conditions!)

1. The _____ adults were _____ selected.
2. It is _____ to believe that 2000 adults is _____ 10% of all adults.
3. All expected counts are _____!

The conditions have all been _____.

	No High School Diploma	High School Diploma, No College	High school Diploma, Some College
Employed	206 (213.4)	548 (552.9)	1186 (1173.7)
Unemployed	14 (6.6)	22 (17.1)	24 (36.3)

(expected counts)

What Should We Take Away?

How do we verify the conditions for performing a chi-square test for homogeneity or independence?

1. **Data should be collected using a _____ random sample or randomized experiment (_____) or a _____ random sample (_____)**
2. **When sampling _____ replacement, the sample size is _____ to 10% of the population size.**
3. **All expected counts should be _____.**

Name_____

AP Statistics CED 8.6 Daily Video 1 (Skill 3.E)

Carrying Out a Chi-Square Test for Homogeneity or Independence

What Will We Learn?

How do we calculate an appropriate test statistic for a chi-square test for homogeneity or independence?

How do we calculate a *p*-value for a chi-square test for homogeneity or independence?

Where Do You Go To School?

A random sample of parents with school-aged children was taken during 2019. A separate random sample of parents with school-aged children was taken in 2020. Parents were asked what type of school their children attended. Here are the results. Researchers would like to perform a test to determine if the distribution of school type for all school-aged children differed between the two years. Find the test statistic and *p*-value.

	2019 Sample	2020 Sample
Public	266	163
Private/Parochial/Charter	16	21
Home	38	30

From Previous Videos

H_0: There is no difference in the distribution of school types attended by school-aged children from 2019 to 2020.

H_a: There is a difference in the distribution of school types attended by school-aged children from 2019 to 2020.

We identified the procedure as a *chi-square test for homogeneity.*

We determined all three conditions are met.

Calculating a Test Statistic

$$expected\ count = \frac{(row\ total)(column\ total)}{table\ total}$$

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

	2019 Sample	2020 Sample	Total
Public	266 (257.1)	163 (171.9)	429
Private/Parochial/Charter	16 (22.2)	21 (14.8)	37
Home	38 (40.7)	30 (27.3)	68
Total	320	214	534

(expected counts)

$$\chi^2 = \frac{(266 - \quad)^2}{\quad} + \frac{(16 - \quad)^2}{\quad} + \frac{(38 - \quad)^2}{\quad} + \frac{(163 - \quad)^2}{\quad} + \frac{(21 - \quad)^2}{\quad} + \frac{(30 - \quad)^2}{\quad}$$

χ^2 = _____ + _____ + _____ + _____ + _____ + _____ ← contributions

χ^2 = _____ ← chi-square statistic

Calculating a *p*-value

The *p*-value is the _____ of observing a chi-square statistic at _____ as large as the one observed, assuming the _____ hypothesis and probability model are _____.

For a chi-square test for homogeneity or independence, the *p*-value is calculated from a chi-square distribution with degrees of freedom given by the formula:

df = (_____ - 1)(_____ - 1)

df = (_____)(_____)

df = _____

	2019 Sample	2020 Sample	Total
Public	266	163	429
Private/Parochial/Charter	16	21	37
Home	38	30	68
Total	320	214	534

Name_____

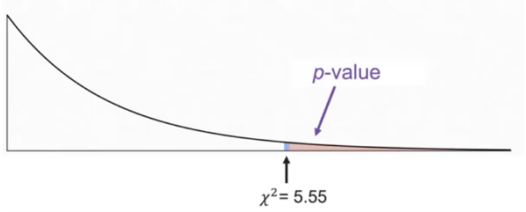
Calculating a p-*value*

$\chi^2 = 5.55$ with df = 2

p-value = $P(\chi^2 \geq 5.55)$ purely by chance!

There are two ways to calculate this *p*-value:

1. Using Table C
2. Using Technology



Calculating *p*-value Using Table C

Using Table C we can only get an estimation

1. Identify the df
2. locate the chi-square statistic
3. find the interval for the *p*-value

Here we can determine that the *p*-value is between _____ and _____.

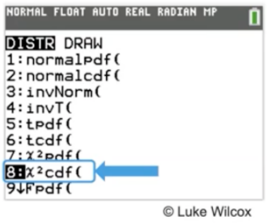
Table C χ^2 **Critical Values**

	Tail probability *p*										
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32			9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74			11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11

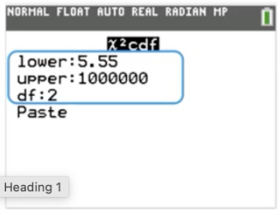
$\chi^2 = 5.55$

Calculating *p*-value With Technology

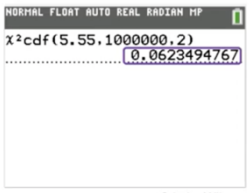
Using technology we can find the exact *p*-value. $P(\chi^2 \geq 5.55)$ = _____



© Luke Wilcox



© Luke Wilcox



© Luke Wilcox

Are you Working?

A random sample of 2000 adults revealed the following data about education level and employment status. Researchers want to determine if there is an association between education level and employment status. Find the test statistic and the *p*-value.

	No High School Diploma	High School Diploma, No College	High School Diploma, Some College
Employed	206	548	1186
Unemployed	14	22	24

From Previous Video

H_0: There is no association between education level and employment status for all adults.

H_a: There is an association between education level and employment status for all adults.

We identified the procedure as a *chi-square test for independence.*

We determined all three conditions are met.

Calculating a Test Statistic

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

	No High School Diploma	High School Diploma, No College	High School Diploma, Some College	Total
Employed	206 (213.4)	548 (552.9)	1186 (1173.7)	1940
Unemployed	14 (6.6)	22 (17.1)	24 (36.3)	60
Total	220	570	1210	2000

(expected counts)

$$\chi^2 = \frac{(206 - \quad)^2}{\quad} + \frac{(548 - \quad)^2}{\quad} + \frac{(1186 - \quad)^2}{\quad} + \frac{(14 - \quad)^2}{\quad} + \frac{(22 - \quad)^2}{\quad} + \frac{(24 - \quad)^2}{\quad}$$

χ^2 = _____ + _____ + _____ + _____ + _____ + _____ ← contributions

χ^2 = _____ ← chi-square statistic

Name_____

Calculating the *p*-value

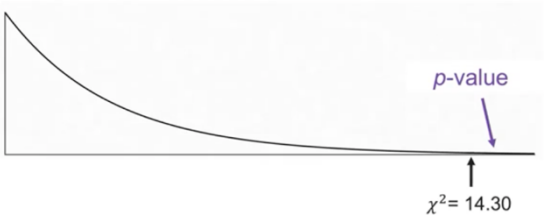
$\chi^2 = 14.30$

df = (_____)(_____) = (_____)(_____) = _____

p-value = $P(\chi^2 \geq 14.30)$

There are two ways to calculate this *p*-value:

1. Using Table C
2. Using Technology



Calculating *p*-value Using Table C

Using Table C we can only get an estimation

1. Identify the df
2. locate the chi-square statistic
3. find the interval for the *p*-value

Table C χ^2 **Critical Values**

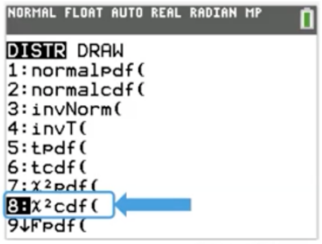
	Tail probability *p*											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11

$\chi^2 = 14.30$

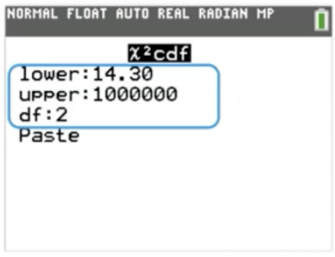
Here we can determine that the *p*-value is between _____ and _____.

Calculating *p*-value Using Technology

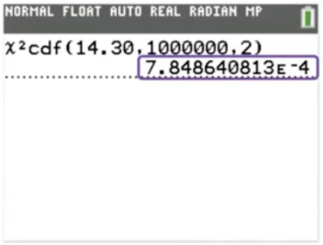
Using technology we can find the exact *p*-value. $P(\chi^2 \geq 14.30)$ = _____



© Luke Wilcox



© Luke Wilcox



© Luke Wilcox

What Should We Take Away?

How do we calculate an appropriate test statistic for a chi-square test for homogeneity or independence?

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

How do we calculate a *p*-value for a chi-square test for homogeneity or independence?

$P(\chi^2 \geq$ _____ test statistic)

df = (_____)(_____)

Use Table ____ or χ^2cdf on the calculator.

STATS MEDIC

Name_____

AP Statistics CED 8.6 Daily Video 2 (Skill 4.B)

Carrying Out a Chi-Square Test for Homogeneity or Independence

What Will We Learn?

How do we interpret the *p*-value in a chi-square test for homogeneity or independence?
How do we state a conclusion for a chi-square test for homogeneity or independence?

Where Do You Go To School?

A random sample of parents with school-aged children was taken during 2019. A separate random sample of parents with school-aged children was taken in 2020. Parents were asked what type of school their children attended. Here are the results. Researchers would like to perform a test to determine if the distribution of school type for all school-aged children differed between the two years. Find the test statistic and *p*-value.

	2019 Sample	2020 Sample
Public	266	163
Private/Parochial/Charter	16	21
Home	38	30

From Previous Videos

H_0: There is no difference in the distribution of school types attended by school-aged children from 2019 to 2020.
H_a: There is a difference in the distribution of school types attended by school-aged children from 2019 to 2020.
We identified the procedure as a *chi-square test for homogeneity.*
We determined all three conditions are met.
$\chi^2 = 5.55$, df = 2, *p*-value = 0.062

The *p*-value measure how _____ it is to get evidence for H_a as _____ as or _____ than the observed evidence _____ when H_0 is _____.

Interpreting a *p*-value

Assuming _____, there is a <*p-value*> _____ of getting a χ^2 of <*calculated chi-square*> or greater, by _____ in the random sample(s) or random assignment.
In context:
Assuming there is _____ in the distribution of _____ for school-aged children from 2019 to 2020, there is a _____ of getting χ^2 = _____ or _____ by chance alone in the random samples.

Stating a Conclusion

Small *p*-values ⟶ test statistic is _____ to occur by random _____ alone.

- Because the *p*-value of _____ ≤ α = _____, we reject H_0.
  There is convincing _____ evidence that [_____].

Large *p*-values ⟶ test statistic is _____ to occur by random _____ alone.

- Because the *p*-value of _____ ≤ α = _____, we fail to reject H_0.
  There is not convincing _____ evidence that [_____].

Stating a Conclusion

No significance level was stated in the school type example, so we will use α = _____, which is the most _____ significance level.
Because the *p*-value of _____, we _____.
There is _____ that there is a _____ in the distribution of school types for school-aged children from _____,

STATS MEDIC

Name_____

Are you Working?

A random sample of 2000 adults revealed the following data about education level and employment status. Researchers want to determine if there is an association between education level and employment status. The *p*-value for a chi-square test for independence is 0.0008. Interpret the *p*-value and make a conclusion at the $\alpha = 0.01$ significance level.

	No High School Diploma	High School Diploma, No College	High School Diploma, Some College
Employed	206	548	1186
Unemployed	14	22	24

From Previous Videos

H_0: There is no association between education level and employment status for all adults.
H_a: There is an association between education level and employment status for all adults.
We identified the procedure as a *chi-square test for independence.*
We determined all three conditions are met. $\chi^2 = 14.30$, df = 2, *p*-value = 0.0008
Here is what the *p*-value actually means:
Assuming there is _____ between educational level and employment status for all adults, there is a 0.0008 _____ of getting a χ^2 of _____ or _____ by chance alone in a random sample of _____.

State A Conclusion

Because the *p*-value of _____, we _____.
There is _____ that the _____ between educational level and employment status for all adults.

Follow-Up Analysis

The largest contribution to the chi-square statistic is _____, because the _____ number of unemployed adults with no high school diploma is much _____ than _____.

	No High School Diploma	High School Diploma, No College	High School Diploma, Some College	Total
Employed	206 (213.4)	548 (552.9)	1186 (1173.7)	1940
Unemployed	14 (6.6)	22 (17.1)	24 (36.3)	60
Total	220	570	1210	2000

(expected counts)

$$\chi^2 = \Sigma \frac{(observed - expected)^2}{expected}$$

$$\chi^2 = \frac{(206-213.4)^2}{213.4} + \frac{(548-552.9)^2}{552.9} + \frac{(1186-1173.7)^2}{1173.7} + \frac{(14-6.6)^2}{6.6} + \frac{(22-17.1)^2}{17.1} + \frac{(24-36.3)^2}{36.3}$$

$\chi^2 = 0.26 + 0.04 + 0.13 + 8.30 + 1.40 + 4.17$ ← contributions

$\chi^2 = 14.30$ ← chi-square statistic

What Should We Take Away?

How do we interpret the *p*-value in a chi-square test for homogeneity or independence?

Assuming _____, there is a <*p-value*> _____ of getting a χ^2 of <*calculated chi-square*> or greater, by _____ in the random sample(s) or random assignment.

How do we state a conclusion for a chi-square test for homogeneity or independence?

- **Because the *p*-value of _____ $\leq \alpha =$ _____, we reject H_0.**
  There is convincing _____ evidence that [_____].
- **Because the *p*-value of _____ $\leq \alpha =$ _____, we fail to reject H_0.**
  There is not convincing _____ evidence that [_____].

STATS MEDIC

Name_____

AP Statistics CED 8.6 Daily Video 3 (Skill 4.E)

Carrying Out a Chi-Square Test for Homogeneity or Independence

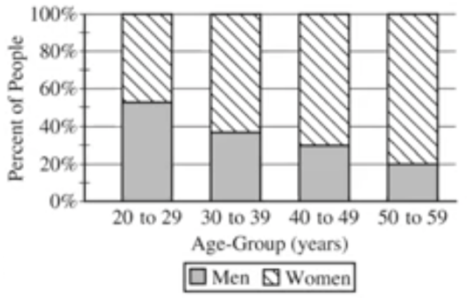
What Will We Learn?

How do we perform a complete chi-square test for homogeneity or independence?

2017 #5

The table and bar chart below summarize the age at diagnosis, in years, for a random sample of 207 mean and women currently being treated for schizophrenia.

	Age-Group (years)				
	20 to 29	30 to 39	40 to 49	50 to 59	Total
Women	46	40	21	12	119
Men	53	23	9	3	88
Total	99	63	30	15	207



Do the data provide convincing statistical evidence of an association between age-group and gender in the diagnosis of schizophrenia?

In the wording of the question there is a clue to us that we need to do a full significance test. Highlight that wording. What type of test do we want here? We have a _____ and we know that we are trying to see if there are two _____ that are associated so we know that this is going to require a *chi-square test for* _____.

2017 #5 Hypotheses

H_0: _____ and _____ are _____ (that is, they are not _____ for the population of people _____ for schizophrenia.

H_a: _____ and _____ are _____ (that is, they are associated) for the population of people _____ for schizophrenia.

No significance level was stated, so we'll use α = _____.

2017 #5 Procedure and Conditions

Procedure: *chi-square test for* _____

Conditions: (Be sure to ✓ the conditions!)

1. _____ men and women currently being treated for schizophrenia.
2. _____ is less that _____ of _____ currently being treated for schizophrenia.
3. All _____ counts are greater than ____.

$$expected\ count = \frac{(row\ total)(column\ total)}{table\ total}$$

Use the formula to calculate all of the expected counts.

	Age-Group (years)				
	20 to 29	30 to 39	40 to 49	50-59	Total
Women	46	40	21	12	119
Men	53	23	9	3	88
Total	99	63	30	15	207

Key: (expected counts)

All conditions have been _____.

Name_____

2017 #5 Test Statistic and *p*-value

		Age-Group (years)			
	20 to 29	30 to 39	40 to 49	50 to 59	Total
Women	46 (56.91)	40 (36.22)	21 (17.25)	12 (8.62)	119
Men	53 (42.09)	23 (26.78)	9 (12.75)	3 (6.38)	88
Total	99	63	30	15	207

(expected counts)

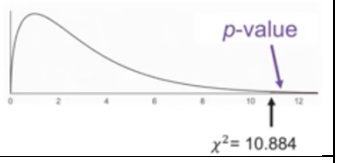
$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Use the formula to calculate the χ^2 statistic.

χ^2 =

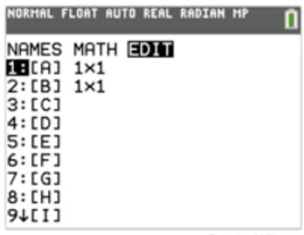
χ^2 =

χ^2 = df = *p*-value =



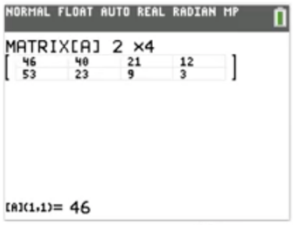
2017 #5 Calculator

Select Matrix



© Luke Wilcox

Enter Observed Data

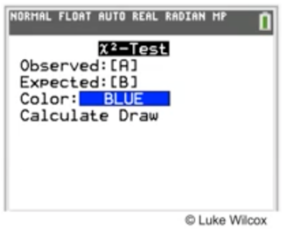


© Luke Wilcox

Select χ^2 Test

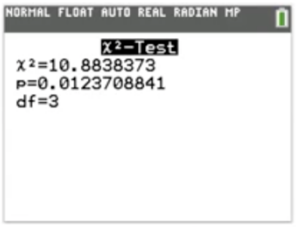
© Luke Wilcox

Select Calculate



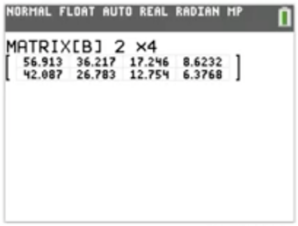
© Luke Wilcox

Test Statistic *p*-value and df



© Luke Wilcox

Expected Counts are now in Matrix B



© Luke Wilcox

2017 #5 Conclusion

Because the *p*-value _____, we _____.

There _____ that there is an _____ between _____ and _____ for the population currently being treated for schizophrenia.

What Should We Take Away?

How do we perform a complete chi-square test for homogeneity or independence?

Make sure to:

- **State the _____ and _____ hypotheses**
- **Identify the _____. Use _____ is not is stated.**
- **Identify the _____ you are using.**
- **Verify that the _____ for the procedure are met (with evidence!).**
- **Calculate the _____ and the _____.**
- **Make a _____ based on the *p*-value. (You do not need to _____ the *p*-value unless specifically asked.)**

STATS MEDIC

Name_____

AP Statistics CED 8.7 Daily Video 1

Skill Focus – Inference Procedures for Categorical Data

What Will We Learn?

How do we identify an appropriate chi-square test for a set of categorical data?

Identifying the Procedure

Goal: to _____ a distribution of a categorical variable to a _____ distribution

Data collection: _____ random sample from _____ population

→ (1) *chi-square test for* _____ **(_____ sample, ____ categorical variable)**

Goal: to _____ distributions of a categorical variable for _____ populations

Data collection: _____ random samples from _____ populations (or from multiple groups in a randomized experiment)

→ (2) *chi-square test for* _____ **(_____ samples, ____ categorical variable)**

Goal: to determine whether _____ categorical variables are _____.

Data collection: _____ random sample from _____ population.

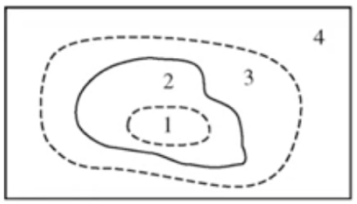
→ (3) *chi-square test for* _____ **(_____ sample, ____ categorical variables)**

2008 #5

A study was conducted to determine where moose are found in a region containing a large burned area. A map of the study area was partitioned into the following four habitat types.

(1) Inside the burned are, not near the edge of the burned area.
(2) Inside the burned area, near the edge.
(3) Outside the burned area, near the edge, and
(4) Outside the burned area, not near the edge.

The figure shows these four habitat types.



Note: Figure not drawn to scale.

2008 #5

The proportion of total acreage in each of the habitat types was determined for the study area. Using an aerial survey, moose locations were observed and classified into one of the of the four habitat types. The results are given in the table below. The researcher who are conducting the study expect the number of moose observed in a habitat type to be proportional to the amount of acreage of that type of habitat. Are the data consistent with this expectation? Conduct an appropriate statistical test to support your conclusion.

Habitat Type	Proportion of Total Acreage	Number of Moose Observed
1	0.340	25
2	0.101	22
3	0.104	30
4	0.455	40
Total	1.000	117

2008 #5

What type of test is appropriate here?

We need to know:

Number of samples: _____

Number of categorical variables: _____

We want compare the distribution of habitat type from the _____ to the _____ distribution that it would be proportional to the amount of acreage.

So, in this case we will use a → *chi-square test for* _____

Name_____

2016 #2

Product advertisers studied the effects of television ads on children's choices for two new snacks. The advertisers used two 30-second television ads in tan experiment. One ad was for a new sugary snack called Choco-Zuties, and the other was for a new healthy snack called Apple-Zuties. For the experiment, 75 children were randomly assigned to one of three groups, A, B, or C. Each child individually watched a 30-minute television program that was interrupted for 5 minutes of advertising. The advertising was the same for each group with the following exceptions.

- The advertising for group A included the Choco-Zuties ad but not the Apple-Zuties ad.
- The advertising for group B included the Apple-Zuties ad but not the Choco-Zuties ad.
- The advertising for group C included neither the Choco-Zuties ad not the Apple-Zuties ad.

2016 #2

After the program, the children were offered a choice between the two snacks. The table below summarizes their choices. Do the data provide convincing statistical evidence that there is an association between type of ad and children's choice of snack among all children similar to those who participated in the experiment?

Group	Type of Ad	Number Who Chose Choco-Zuties	Number Who Chose Apple-Zuties
A	Choco-Zuties only	21	4
B	Apple-Zuties only	13	12
C	Neither	22	3

2016 #2

What type of test is appropriate here? Since this is a randomized experiment, we need to know:

Number of groups: _____

Number of categorical variables: _____

So, in this case we will use a → *chi-square test for* _____

2013 #4

The Behavioral Risk Factor Surveillance System is an ongoing health survey system that tracks health conditions and risk behaviors in the United States. In one of their studies, a random sample of 8,866 adults answered the question "Do you consume five or more servings of fruits and vegetables per day?" The data are summarized by response and by age-group in the frequency table below. Do the data provide convincing statistical evidence that there is an association between age-group and whether or not a person consumes five or more servings of fruits and vegetables per day for adults in the United States?

Age-Group (years)	Yes	No	Total
18–34	231	741	972
35–54	669	2,242	2,911
55 or older	1,291	3,692	4,983
Total	2,191	6,675	8,866

2013 #4

What type of test is appropriate here? Since this is a randomized experiment, we need to know:

Number of samples: _____

Number of categorical variables: _____

So, in this case we will use a → *chi-square test for* _____

What Should We Take Away?

How do we identify an appropriate chi-square test for a set of categorical data?

___ sample, _____ categorical variable → (1) *chi-square test for* _____

___ samples, _____ categorical variable → (1) *chi-square test for* _____

___ sample, _____ categorical variables → (1) *chi-square test for* _____