

Chapter 23. Inference on Multiple Proportions

1. What does it mean to move from summarizing two-variable categorical data (Ch. 6) to doing inference on the data (Ch. 23)?
2. How are these problems different from comparing two proportions in Ch. 21?
3. What are the calculations we need to do and how do we do them?
 - a. Software
 - b. Understanding / By hand:
 - i. What does it take to describe what the data would look like if H_0 is true?
 - ii. How do we summarize how different the actual data is from what it would be if H_0 is true?
 - iii. What if a person actually did all the computations for this with a calculator so the software didn't give the p-value. How do we use a table to find it?
 - iv. How many of these computations do we have to do by hand / calculator?
4. How large does our sample size have to be for these procedures to be used?
5. What is the section "Uses of the Chi-Square Test" about?
6. Can this procedure be used in situations where both variables have only two possible outcomes (so we could also have used the Ch. 21 methods)?
Answer: Yes.
7. Do we cover the last section of this Chapter in this course? (The Chi-Square Test for Goodness of Fit)
Answer: No

Discussion:

1. What does it mean to move from summarizing two-variable categorical data (Ch. 6) to doing inference on the data (Ch. 23)?
Answer: We could summarize a two-way table by calculating various distributions. But how could we tell whether two such distributions were different enough to be meaningful, or could such a difference have happened by chance if the population dist'ns were not different? We couldn't back then. Now we know about inference and need to explore how to do it here. **See the overall summary on page 8 here.**
2. How are these problems different from comparing two proportions in Ch. 21?
Answer: We often have more than two proportions to compare. We want a method to see if all the proportions are the same or whether some are different.
8. What are the calculations we need to do and how do we do them?
Answer: For this chapter, it is best to have one overall activity to illustrate the method. We will do various pieces of it and summarize what they mean.

Relationships: Categorical Variables

- Chapter 21: compare proportions of successes for two groups
 - “Group” is explanatory variable (2 groups)
 - “Success or Failure” is response (2 outcomes)

- Chapter 23: “*Is there a relationship between two categorical variables?*”
 - may have 2 or more groups (one variable)
 - may have 2 or more outcomes (2nd variable)

- Recall from Chapter 6:
 - When there are two categorical variables, the data are summarized in a **two-way table**
 - The number of observations falling into each combination of the two categorical variables is entered into each **cell** of the table
 - Relationships between categorical variables are described by calculating appropriate **percents** from the counts given in the table

Students and catalog shopping. What is the most important reason that students buy from catalogs? The answer may differ for different groups of students. Here are results for samples of American and East Asian students at a large Midwestern university:				Activity 1: What percentages should you compute to see if the different groups behave differently in terms of their reasons for catalog shopping? Compute several of the percentages. (Hint: Conditional dist’ns of the response variable.)			
Reason	American	Asian	Totals	Reason	American	Asian	Totals
Save time	30	10	40	Save time			40
Easy	29	20	49	Easy			49
Low Price	18	36	54	Low Price			54
Live Far from Stores	11	4	15	Live Far from Stores			15
No pressure to buy	10	3	13	No pressure to buy			13
Other	22	7	29	Other			29
Totals	120	80	200	Totals	120	80	200

From Chapter 6, if the conditional distributions of the response variable are nearly the same for each group, then we say that there is not an association between the two variables. It was frustrating in Ch. 6 to not be able to say how different those conditional distributions had to be for us to say there were *big enough differences* to claim that these groups are “significantly different” in their distributions. So we really want to perform some kind of hypothesis test.

Activity 2: In the blank space at the end of this page, write the hypotheses here for THIS hypothesis test.

Hypothesis Test (Test of Significance)

- In tests for two categorical variables, we are interested in whether a relationship observed in a sample reflects a real relationship in the population.
- Hypotheses: (Today let's test at the 5% significance level.)
 - H_0 : **There is no real relationship between _____ and _____**
(The percentages for one variable are the same for every level of the other variable in the population(s).)
(There is no difference in conditional distributions in the population(s).)
 - H_a : **There is a real relationship between _____ and _____**
(The percentages for one variable vary over levels of the other variable in the population(s).)
(There is a difference in the conditional distributions in the population(s)).

Expected Counts

If H_0 is true – if there is really no relationship between “reason” and “group” here, then the same proportions we see in the totals should be reflected in all the rows.

Reason	American	Asian	Totals
Save time			40
Easy			49
Low Price			54
Live Far from Stores			15
No pressure to buy			13
Other			29
Totals	120	80	200

That is – the Americans form $120/200 = 60\%$ of the overall numbers, so they should be 60% of the people who give “save time” as their reason. So that means the expected value in that first cell should be 60% of 40, which is 24.

Activity 3a: Fill in that in the first “cell” and then continue that same idea to fill in the expected counts (rounded to two decimal places) for **only the first two rows**. Do that on the previous page if you are confident you know what to do. Do it in the table just below if you need help.

Notice that we found each of these expected counts by

Reason	American	Asian	Totals
Save time	$\frac{120}{200} \cdot 40 = \frac{120 \cdot 40}{200} = 24$	$\frac{80}{200} \cdot 40 = \frac{80 \cdot 40}{200} = 16$	40
Easy			49

This suggests the formula that we usually use for the expected count in any cell of a two-way table

(when H_0 is true):
$$\text{expected count} = \frac{\text{column total} \times \text{row total}}{\text{table total}}$$

Activity 3b: Fill in **at least three** of the rest of the expected counts using this formula.

$$\text{expected count} = \frac{\text{column total} \times \text{row total}}{\text{table total}}$$

As far as you can, with what you have done, check to see that your totals agree with the totals given.

Reason	American	Asian	Totals
Save time			40
Easy			49
Low Price			54
Live Far from Stores			15
No pressure to buy			13
Other			29
Totals	120	80	200

The Chi-Square Test Statistic

- To determine if the differences between the observed counts and expected counts are statistically significant (to show a real relationship between the two categorical variables), we use the **chi-square statistic**: $\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$ where we add up this calculation for each cell in our table.
- The chi-square statistic is a measure of the distance of the observed counts from the expected counts
- is always zero or positive
 - is only zero when the observed counts are exactly equal to the expected counts
 - large values of χ^2 are evidence against H_0 because these would show that the observed counts are far from what would be expected if H_0 were true
 - the chi-square test is one-sided (any violation of H_0 produces a large value of χ^2)

For our example:

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} = \text{Sum of twelve terms}$$

$$\chi^2 = \frac{(30-24)^2}{24} + \frac{(10-16)^2}{16} + \frac{(29-29.40)^2}{29.40} + \dots$$

$$\chi^2 = 1.500 + 2.250 + 0.005 + \dots$$

Activity 4: You fill in two more terms of the chi-squared statistic, showing your work, in the lines above. Use the original data (below) and the expected counts you computed on the previous page.

Recall that these were the original data values:

Activity 5: Do the “Solve” part of this problem using Minitab. To obtain this Minitab output, type in the table (without totals) into the worksheet. Put the labels “American” and “Asian” in the row above the data rows. You can put in the labels for the reasons, but the output won’t show it. Make sure your output agrees with the output given here. (It won’t if you put the Total row in!)

Stat > Tables > Chi-squared test (Table in worksheet)

Chi-Square Test: American, Asian

Expected counts are printed below observed counts
 Chi-Square contributions are printed below expected counts

	American	Asian	Total
1	30	10	40
	24.00	16.00	
	1.500	2.250	
2	29	20	49
	29.40	19.60	
	0.005	0.008	
3	18	36	54
	32.40	21.60	
	6.400	9.600	
4	11	4	15
	9.00	6.00	
	0.444	0.667	
5	10	3	13
	7.80	5.20	
	0.621	0.931	
6	22	7	29
	17.40	11.60	
	1.216	1.824	
Total	120	80	200

Chi-Sq = 25.466, DF = 5, P-Value = 0.000

We could check this by adding all the contributions to the Chi-Sq and see that they do sum to 25.466.

Activity 6: Read the P-value from the software and write a conclusion “in context”, recalling that we said we’d use the 5% significance level.

To find the P-value by hand, we have to learn to read a table for a new distribution, in a later section in this chapter.

Chi-Square Test Conditions

- The numbers we use in the chi-square test must be COUNTS, not percentages.
- The data can be an SRS from a single sample where each individual is classified on two variables OR from independent SRSs from two or more populations with each individual classified on one categorical variable.
- Each individual represented in the table must show up in ONLY ONE cell of the table. (That’s implied by the previous condition, but is very important, so I re-state it.)
- The chi-square test is an approximate method, and is more accurate if the counts in the cells of the table get larger (actually, it is the EXPECTED counts that need to be large enough.)
- The following must be satisfied for the approximation to be accurate:
 - No more than 20% of the expected counts are less than 5
 - All individual expected counts are 1 or greater
- If these conditions fail, then two or more groups must be combined to form a new (‘smaller’) two-way table

Chi-Square Test

- Calculate value of chi-square statistic
 - by hand (cumbersome)
 - using technology (computer software, etc.)
- Find P -value in order to reject or fail to reject H_0
 - use chi-square table for chi-square distribution (later in this chapter)
 - from computer output
- If significant relationship exists (small P -value):
 - look at individual terms in the chi-square statistic (easy to do) AND
 - compare individual observed and expected cell counts (easy to do) OR
 - compare appropriate percents in data table (more tedious to do)

Summary: In this handout, you had these activities.

1. Find the conditional distributions to answer the question about whether the different groups differed in their online shopping reasons.
2. Write the hypotheses in words (Fill in some blanks.)
3. Find the expected values if H_0 is true. (a few)
4. Find each cell's contribution to the chi-squared statistic. (a few)
5. Produce the Minitab output to do this test.
6. Write a conclusion based on the p-value produced by the software.

Still to do:

7. If a significant relationship exists, use data analysis to discuss its direction.
8. Find the appropriate table, compute the degrees of freedom, and find the P-value (an interval for it) with the table instead of with software. (See the section in the book "The chi-squared distributions." p. 570.)
9. Since the chi-squared table is NOT centered at zero, you need a "reference value" to get a rough idea of what a particularly large chi-squared value is. Answer: The center of a chi-squared dist'n is equal to its degrees of freedom.
10. Know that we DO NOT cover the last section in the text "Chi-squared test for goodness of fit."
11. Discuss and practice checking the conditions for a chi-squared test.

Activity 7:

- a. Do exercise 23.7 in the section "Using Technology." Use a 1% significance level. Note that you do not have to actually input the data into Minitab – you can use the given Minitab output.
- b. For the "data analysis" part of your solution, do not compute conditional distributions, but follow the model of Example 23.5 in the section "Using Technology" on page 564.
- c. After you have done the problem as stated, go back and practice by doing these:
 - i. Compute at least two expected values "by hand" and show your work.
 - ii. Compute the contributions to the chi-squared statistic from those two cells "by hand" and show what you are computing.
 - iii. Find the degrees of freedom for this problem.
 - iv. Use the table to find the P-value for this problem.

Activity 8: Do exercises 23.36 – 23.39 and discuss your answers with others in your group.

Activity 9: Test these hypotheses with EACH of these two datasets. Use your results to explain why it is not correct to do chi-squared tests when the entries in each cell are percentages instead of counts.

Ho: There is no relationship between a person's sex and their political preference.

Ha: There is a relationship between a person's sex and their political preference.

Sample A				Sample B			
	Male	Female	Total		Male	Female	Total
Democrat	4	6	10	Dem	4000	6000	10000
Republican	6	4	10	Repub.	6000	4000	10000
Total	10	10	20		10000	10000	20000

Quiz 13: Due Wed. April 24, at the beginning of class. Submit a word-processing document in this quiz assignment in Blackboard with your solutions for this quiz, as if it were a test question. Include your answers in the main part and your computer output in the Appendix. In the Appendix, identify the output for each of the individual problems separately. You have ONE opportunity to submit this, just as you will for the test.

First problem: (30 points) Video games and grades: For the data discussed in exercises 23.2, 23.4, and 23.8, is the evidence of a relationship between grades and playing video games. You do not have to follow the instructions in any of those problems. Instead, solve this using the four-step process, including, if there is significant evidence of a relationship, do a data analysis to discuss the direction.

Second problem: (20 points) do 23.36 and 23.38

Third problem: (50 points) 23.28.