

Lab 9: Categorical Data

Objectives:

1. Understand how to analyze categorical data
2. Understand how to perform chi-square tests in R.

Definitions:

- categorical (qualitative) data
- chi-square distribution
- observed vs. expected counts
- goodness-of-fit test
- contingency table
- test of homogeneity
- test of independence

Introduction:

Recall that *categorical data* is data based on some attribute or characteristic. The observations fall into *categories*. Up to this point, we have performed hypothesis tests primarily about population means. But if we are interested in testing claims about categorical data, then we need a new approach, since we cannot compute means for categorical variables. Instead we focus on proportions, and we have only developed tests for comparing two proportions at a time. In this lab, we will look at methods to analyze relationships between categorical variables and to check how well a probability model fits a single categorical variable.

Activities:

Getting Organized: *If you are already organized, and remember the basic protocol from previous labs, you can skip this section.*

Navigate to your class folder structure. Within your "Labs" folder make a subfolder called "Lab9". Next, download the lab notebook .Rmd file for this lab from Blackboard and save it in your "Lab9" folder. There are no datasets used in this lab.

Within RStudio, navigate to your "Lab9" folder via the file browser in the lower right pane and then click "More > Set as working directory". Get set to write your observations and R commands in an R Markdown file by opening the "lab9_notebook.Rmd" file in RStudio. Remember to add your name as the author in line 3 of the document. For this lab, enter all of your commands into code chunks in the lab notebook. You can still experiment with code in an R script, if you want. To set up an R Script in RStudio, in the upper left corner click "File > New File > R script". A new tab should open up in the upper left pane of RStudio.

Goodness-of-Fit Tests: In class on Tuesday, we considered whether any one day of the week is more or less likely to be a person's birthday than any other day of the week. Let p_M denote the proportion of *all* people that were born on a Monday, or equivalently, the probability that a randomly selected person was born on a Monday. Similarly, define p_{Tu} , p_W , p_{Th} , p_F , p_{Sa} , and p_{Su} . We are testing the following hypotheses:

$$H_0 : p_M = p_{Tu} = p_W = p_{Th} = p_F = p_{Sa} = p_{Su} = 1/7$$

$$H_A : p_i \neq 1/7 \text{ for at least one day of the week}$$

In other words, we are testing whether the probability model stated in the null hypothesis fits the data well.

To test these hypotheses, you created a version of the following table:

	Days of the Week					Total: n
	Mon	Tues	Wed	Thu		
		Fri	Sat	Sun		

Observed counts: O_i	17	26	22	23	147
	19	15	25		
Expected counts: $E_i = np_i$	21	21	21	21	147
	21	21	21		
$(O_i - E_i)^2 / E_i$	0.76	1.19	0.05	0.19	4.86
	0.19	1.71	0.76		

The test statistic in this case, 4.86, follows a [chi-square distribution](#), with degrees of freedom equal to the number of categories (i.e., days of the week) minus one, and so the P -value is calculated in R as follows:

```
pchisq(4.86, df = 6, lower.tail = FALSE)
```

```
## [1] 0.5618907
```

Pause for Reflection #1:

Suppose we suspect that weekend days are less likely to be a birthday, perhaps because doctors want the weekend off and so do not schedule Caesarean deliveries for weekends. Let's test whether the data provide evidence against the hypothesis that weekend days are half as likely as other days to be someone's birthday and that all weekdays are equally likely.

- State the hypotheses being tested in this case. The null hypothesis should give the proposed probability model for the data. Note that not all the days of the week will have the same probabilities, but we will still need the probabilities to add up to 1.
- Redo the table above to calculate the test statistic in this case. Note that we are using the same data, so the observed counts stay the same, but the expected counts will change.
- Alter the R code above to calculate the corresponding P -value and state the conclusion of the test.
- Which category (day) has the largest contribution to the test statistic? Explain what this reveals.

Chi-Square Test in R: As you may have already guessed, there is a function in R, `chisq.test()`, that performs the calculations you just did. To use this function, store the observed counts in a list:

```
birthdays = c(17, 26, 22, 23, 19, 15, 25)
```

For the test that each day of the week is equally likely, all we have to do is call the `chisq.test()` function on the object containing the observed counts as follows, since by default R tests the data against the null hypothesis that all probabilities are equal:

```
chisq.test(birthdays)

##
## Chi-squared test for given probabilities
##
## data: birthdays
## X-squared = 4.8571, df = 6, p-value = 0.5623
```

The output above gives the value of the observed test statistic χ^2 and the degrees of freedom df for the chi-square distribution used to calculate the corresponding p -value.

For the test that weekend days are half as likely as other days, we need to specify the probabilities stated in the null hypothesis in the `chisq.test()` function as follows:

```
probs = c(rep(1/6, 5), 1/12, 1/12)
chisq.test(birthdays, p = probs)
```

Pause for Reflection #2:

Explain the code above, specifically the line defining the object `probs`. Does the output of the `chisq.test` match the results you found in Reflection #1?

Newspaper Reading: Are Americans today less likely to read a newspaper every day than in previous years? The General Social Survey (GSS) interviews a random sample of adult Americans every two years, and one of the questions asks respondents, "How often do you read the newspaper?" Sample results for the years 1978, 1988, 1998, 2008, and 2018 are given in the *contingency table* below.

	1978	1988	
	1998	2008	total
	2018		
Every day	874	500	
	805	431	
	321		2922
Not every day	654	488	
	1065	898	4352
	1247		
	1528	988	
total	1870	1329	7274
	1559		

In asking whether or not these sample data provide evidence that the proportion of Americans who read the newspaper every day differed among the five populations for these years, we have to ask how likely it is to have observed such sample data if, in fact, the "every day" proportions were the same for all five populations (years). However, it's a little harder to quantify this now that we are comparing more than two groups.

We adopt a strategy similar to the goodness-of-fit test: Compare the *observed* counts in the table with the counts *expected* under the null hypothesis of equal population proportions/distributions. The farther the observed counts are from the expected counts, the more extreme we will consider the data to be.

Pause for Reflection #3:

Use appropriate symbols to state the null hypothesis that the population proportion of adult Americans who read the newspaper every day was the same for these five years: 1978, 1988, 1998, 2008, and 2018.

Pause for Reflection #4:

For the five years combined, what proportion of respondents read the newspaper every day? If this same proportion of the 1528 respondents in the year 1978 had read the newspaper every day, how many people would this represent? Record your answer with two decimal places, and repeat for the other four years.

We have now calculated the *expected counts* under the null hypothesis that the population proportion of adult Americans who read the paper every day was the same for these four years (and consequently also the population proportions who did not read the paper every day). A more general technique for calculating the expected count of cell i is to take the marginal total for that row times the marginal total for that column, divided by the grand total (sample size of the study, n):

$$E_i = \frac{\text{row total} \times \text{column total}}{\text{grand total}} \quad (\text{Lab 9.1})$$

Pause for Reflection #5:

Use the general formula in Equation (9.1) to calculate the expected count of "not every day" people in the year 1988 and complete the following table:

	1978 1998 2018	1988 2008	
	874 805 312 (613.80) (751.19) (626.26)	500 431 (396.88) (533.87)	total
Every day			2922
	654 1065 1247 (914.20) (1118.81) (932.74)	488 898 () (795.13)	
Not every day			4352
	1528 1870 1559	988 1329	
total			7274

Now that we have the observed counts and the expected counts calculated, we need to find a *test statistic* to measure how far the observed counts deviate from the expected counts. To do this, we do the same calculation as with the goodness-of-fit test:

$$X^2 = \sum_{\text{all cells } i} \frac{(O_i - E_i)^2}{E_i}$$

Pause for Reflection #6:

Calculate the value of $(O_i - E_i)^2 / E_i$ for the "not every day" people in 1988 (i.e., for the second cell in the second row of the table). Add this value to other contributions to the test statistic calculation provided below and compute the test statistic:

$$X^2 = 110.30 + 26.79 + 3.85 + 19.82 + 157.70 \\ + 74.06 + ?? + 2.59 + 13.31 + 105.88 = ??$$

What kind of values (e.g., large or small) of the test statistic provide evidence against the null hypothesis that the five populations (years) have the same proportion of Americans reading the newspaper every day? Explain.

Again in this case, the test statistic follows a **chi-square distribution**. However, in this case, the degrees of freedom are equal to $(r - 1)(c - 1)$, where r is the number of rows and c is the number of columns in the contingency table.

Pause for Reflection #7:

Calculate the degrees of freedom for the test statistic found in Reflection #6 and then use the `pchisq()` function to find the corresponding P -value. Based on the P -value, state your conclusion.

Tests of Homogeneity: The test we just performed is called a **chi-square test of equal proportions (homogeneity)**. It is used to test whether the proportions for independent samples from three or more populations are the same. And the calculations can also be done in R with the `chisq.test()` function. First, we need to format the observed counts in R, which can be done using the `rbind()` command:

```
years = rbind(c(874, 500, 805, 431, 312), c(654, 488, 1065, 898, 1247))
years

##           [,1]    [,2]    [,3]    [,4]    [,5]
## [1,]      874     500     805     431     312
## [2,]      654     488    1065     898    1247
```

Then, we simply call the `chisq.test()` function on the table of observed counts `years`:

```
chisq.test(years)

##
## Pearson's Chi-squared test
##
## data: years
## X-squared = 532.28, df = 4, p-value < 2.2e-16
```

We can see the expected counts in R with the following code:

```
chisq.test(years)$expected

##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,]    613.8048    396.8842    751.187    533.8655    626.2876
## [2,]    914.1952    591.1158   1118.8122    795.1345    932.7424
```

Tests of Independence: We continue to consider the GSS survey. But this time, we use only the year 2018 with another variable: the respondent's political inclination, classified as liberal, moderate, or conservative. The sample results are summarized in the

table:

	Liberal Conservative	Moderate
	109	153
	160	
Every day	85	109
Few times a week	95	
Once a week	52	82
Less than once a week	63	
Never	56	68
	64	
	52	65
	63	

Notice how this data is different from the data used in the previous example regarding newspaper reading. In this case, we have *one* random sample of individuals (2018 respondents) that are classified according to *two* variables (political inclination and how often they read the newspaper). Previously, we had *five* separate random samples (for the five years) that were classified on just *one* variable.

It turns out that the same chi-square test applies to two-way tables where the data are one random sample from a population classified on two variables. The difference in the null hypothesis being tested is that, in the population, the two variables are *independent*, and the alternative hypothesis is that there is a *relationship* between the variables.

For the above data, we perform a **chi-square test of independence** for the following hypotheses:

H_0 : political inclination and how often someone reads the paper are independent

H_A : political inclination is related to how often someone reads the paper

Pause for Reflection #8:

Format the data in R using the `rbind()` function. Then call the `chisq.test()` function on the data to perform the calculations for the test of independence. Record your conclusion in your lab notebook. If the test indicates strong evidence of a relationship between the variables, examine the table cells that contribute most to the value of the test statistic in order to describe the relationship.

Lab 9: Categorical Data is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.