Lab 7: Inference for Categorical Data

Overview

In this lab, we do an analysis of categorical variables. It includes tests comparing two proportions and the chi-square test of association. We also examine the chi-squared distribution.

The Data: Gardasil vaccination

HPV (human papillomavirus) is a virus that has been linked to the development of cervical cancer, other anogenital cancers, and genital warts. Gardasil, developed by Merck Laboratories, was licensed by the U.S. Food and Drug Administration in 2006. The FDA recommended Gardasil for use by women aged 9-26. The "typical" Gardasil regimen consists of a sequence of three shots, which should be completed within 12 months. Researchers at Johns Hopkins Medical Institutions (JHMI) gathered the data as part of an attempt to characterize young female patients who complete the anti-HPV Gardasil vaccination sequence. The study subjects are females aged 11-26 who (1) made their first "Gardasil visit" to a Johns Hopkins Medical Institution clinic between 2006 and 2008, and (2) had 12 months to complete the regimen. The data set `gardasil.txt` contains ten variables.

Variable	Description
`Age`	Patient's age in years
`AgeGroup`	Patient's age group (0 = 11-17, 1 = 18-26)
`Race`	Patient's race (0 = white, 1 = black, 2 = Hispanic, 3 = other/unknown)
`Shots`	Number of shots completed
`Completed`	Did the patient complete the three-shot sequence within a 12-month period? (0 = no, 1 = yes)
`InsuranceType`	Type of insurance: 0 = medical assistance, 1 = private payer [Blue Cross Blue Shield, Aetna, Cigna, United, Commercial, CareFirst], 2 = hospital based [EHF], 3= military [USFHP, Tricare, MA])
`MedAssist`	Medical assistant indicator variable: 0 = patient does not have medical assistance, 1 = patient has medical assistance)
`Location`	Clinic that patient attended: 1 = Odenton, 2 = White Marsh, 3 = Johns Hopkins Outpatient Center, 4 = Bayview)
`LocationType`	Location type indicator variable (0 = suburban, 1 = urban)
`PracticeType`	Type of practice patient visited (0 = pediatric, 1 = family practice, 2 = OB-GYN)

Getting started

Get started by importing and exploring the data. Note that this data should look familiar to you since we used it in the last lab. Since it is a `.txt` file, we need to use the `read.table` function.

```
#Set working directory
#setwd("~/Your/Working/Directory/")

#Read in the data
```

```
gardasil<-read.table("gardasil.txt",header=TRUE)

#Explore the data
str(gardasil)
summary(gardasil)
```

Two sample z test

Does the completion rate vary by age group? To answer this, we can compare the proportion who completed the 3-dose sequence among the 11-17 year-old age group to those who completed the 3-dose sequence among the 18-26 year-old age group. The hypotheses can be stated as either: H_0 *: p1 = p2* versus H_a *: p1 ≠ p2* or H_0 *: p1 - p2 = 0* versus H_a *: p1 - p2 ≠ 0.* These two sets of hypotheses are interchangeable.

We can test the hypotheses with the two proportion z-test. Begin by examining your descriptive statistics closely so that you can verify that R is testing the intended conditional proportions.

```
# Create frequency table
Age_Completion_Table<-table(gardasil$AgeGroup,gardasil$Completed)

# View table
Age_Completion_Table

#Add summary margins
addmargins(Age_Completion_Table)

#Calculate row proportions
prop.table(Age_Completion_Table,margin=1)
```

Among the 11-17 year-olds, 35.2% completed the sequence compared to 31.2% among the 18-26 year-olds. To test if the difference is statistically significant, you can use the prop.test again by either inputting the counts or the data. The manual input option is handy if you only have summarized data rather than an actual data set.

```
#Two sample proportion test
prop.test(c(247,222),c(701,712),correct=F)
```

The first argument is the number of participants who completed the vaccination sequence in each age group. The second argument is the total number of study participants in each age group. Alternatively, you can input the contingency table for the two variables. However, when using this method, you must be very careful when specifying the table's organization to ensure the test is comparing the correct two proportions. If you do not correctly enter two proportions, the confidence interval for the difference of the two proportions is meaningless. Always review the printed sample proportions to ensure you are doing the correct analysis.

```
#Two sample proportion test (using table, where the RESPONSE of interest is the first column)
prop.test(Age_Completion_Table,correct=F)
```

This table is organized such that the groups of interest are on the rows, and the outcome of interest is in the first column. Note that the results do not match the manual input we did the first time. The test statistic and results are the same, but the confidence interval is incorrect. It is because our variable `Completed` is in the wrong order. Just like last week, we can correct it by doing this:

```r
# Create new variable with correct ordering
gardasil$Completed2 <- factor(gardasil$Completed, levels=c("yes","no"))

# Create new table
Age_Completion_Table2<-table(gardasil$AgeGroup,gardasil$Completed2)
#View table
Age_Completion_Table2

#Two sample proportion test (using table, where the RESPONSE of interest is the first column)
prop.test(Age_Completion_Table2,correct=F)
```

Now, our results exactly match the output from the manual calculation and are more interpretable. We can talk about the percentage of individuals who completed the vaccine (which we care about).

The last two numbers in the output are our sample estimates – the same values we saw in the `prop.table`. We can subtract those to get the estimated difference in the proportion between 11-17 year-olds and 18-26 year-olds that completed the Gardasil vaccination: 0.04. We are 95% confident that the true difference of the two proportions is in the interval -0.01 to 0.09. This confidence interval contains zero, which makes it plausible that the true difference in proportions is zero. The p-value for the two-sample proportion test is 0.11, which is greater than a 0.05 level of significance. Therefore, we fail to reject the null hypothesis and conclude that there is no significant difference in the proportion of 11-17 and 18-26 year-olds who completed the Gardasil vaccine.

Again, this test yields a chi-squared test statistic with one degree of freedom. If you calculate the z test statistic by hand, it would be $z = \sqrt{X^2} = \sqrt{2.62} = 1.62$.

Chi-Square Test

In this second part, we will look at the Chi-Square test. To understand it, we can ask if the completion rate of the vaccination varies by insurance type? To answer this, we have to compare the completion rates for the four groups of insurance types: hospital-based, medical assistance, military, and private payer. However, we cannot use a proportion test because we have more than two groups. So, we will conduct this test using the chi-square test of association. As always, begin by examining the descriptive statistics.

```r
#Table of frequencies
Insurance_Completion_Table<-table(gardasil$InsuranceType,gardasil$Completed)

#View table
Insurance_Completion_Table

#Add summary margins
addmargins(Insurance_Completion_Table)

#Calculate row proportions
prop.table(Insurance_Completion_Table, margin=1)
```

Here, we see that individuals on “medical assistance” insurance have the lowest completion rate (20.0%), and those on “hospital-based” insurance have the highest completion rate (46.4%). Use the command `chisq.test` to perform the chi-square test of association. With this command, you can input the two variables separately or enter the contingency table showing the relationship between the two variables.

```
#Chi-Square Test for completion by insurance type
chisq.test(gardasil$Completed,gardasil$InsuranceType,correct=F)
chisq.test(Insurance_Completion_Table,correct=F)
```

Similar to the prop.test, use `correct=F` to suppress the continuity correction used for data with small counts. An assumption of the chi-square test is that all expected cell counts are at least 5. You can save the chi-squared test results as an object to view the expected cell counts to check this assumption.

```
#Run and save chi square test for completion by insurance type
Ins.Comp.test<-chisq.test(gardasil$Completed,gardasil$InsuranceType,correct=F)

#View expected cell count
Ins.Comp.test$expected
```

All expected cell counts are at least 5, so conditions are satisfied for valid inference.

Fisher's Exact Test

Fisher's exact test is an alternative to the chi-squared test when at least one expected cell count is less than 5. Like the chi-squared test, Fisher's exact test can be implemented by either inputting variable names or a table of summary data.

```
# Run Fisher's Exact with variables
fisher.test(gardasil$Completed,gardasil$InsuranceType)

#Run Fisher's Exact with existing table
fisher.test(Insurance_Completion_Table)
```

When all expected cell counts are at least 5, conclusions from the chi-squared test and Fisher's exact test will not vary greatly; otherwise, results may differ.

Chi-squared distribution

The chi-squared distribution is generally right-skewed. To calculate an area under the curve for the chi-square distribution, use the `pchisq` command. P-values based on a chi-squared test statistic are given by the upper area under the tail of the distribution; by default, `pchisq` returns a lower tail area. To obtain the upper tail, take the complement.

For example, when we examined the relationship between AgeGroup and Completion, we got a chi-squared test statistic of 2.62 on 1 degree of freedom. The p-value for this test of association can be calculated as follows:

```
1-pchisq(2.62,df=1)
```

It gives a result of 0.1055. Recall that we do not need to multiply this value by 2 because chi-square tests are always one-tailed due to the shape of the distribution. A one-tailed chi-square p-value is equivalent to a 2-tailed z test.