Introduction to Statistical Inference (QTM 100 Lab)

Lecture 7: Inference for Categorical Data

Justin Eloriaga — Emory University

Fall 2024

Gameplan

Data Preliminaries

Inference for Categorical Data

For today...

- We will discuss tests covering
  - Comparing Two Proportions
  - χ^2 test of association

Gardasil

Gardasil, developed by Merck Laboratories, was licensed by the U.S. Food and Drug Administration in 2006 to vaccinate against HPV.

- The “typical” Gardasil regimen consists of a sequence of three shots, which should be completed within 12 months.
- The dataset tries to characterize young female patients who complete the anti-HPV Gardasil vaccination sequence.

Importing the Dataset (again)

- Like before, we can use point-and-click or the working directory

```r
setwd("YourFilePath")
```

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

- Let's also examine the structure and give an overview of the dataset

```r
str(gardasil)
```

```r
summary(gardasil)
```

Two-sample z test

Completed and AgeGroup

- Does the completion rate vary by age group?
  - We can compare the proportion who completed the Gardasil vaccine among the 11-17 year-old age group compared to these who completed the Gardasil vaccine among the 18-26 year-old age group.
- We can express the hypothesis as follows

$$H_0 : p_1 = p_2 \qquad H_a : p_1 \neq p_2$$

- Alternatively, we can express this also as

$$H_0 : p_1 - p_2 = 0 \qquad H_a : p_1 - p_2 \neq 0$$

Completed and AgeGroup

We can create a table using the table() command. Then, we use the addmargins() command to be able to add column and row sums. Then, we can calculate the row propotion using the prop.table() command with the margin option set to 1.
Clearly, 35.2% of the 11-17 year olds completed Gardasil while only 31.2% of the 18-26 year olds did the same.

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

        no yes
  11-17 454 247
  18-26 490 222
> # Add summary margines
> addmargins(Age_Completion_Table)

         no  yes  Sum
  11-17  454  247  701
  18-26  490  222  712
  Sum    944  469 1413
> # Calculate the row proportions
> prop.table(Age_Completion_Table, margin = 1)

               no       yes
  11-17 0.6476462 0.3523538
  18-26 0.6882022 0.3117978
```

Running a Two Sample Proportion Test

To test if the difference is statistically significant, we can use the `prop.test()` command.

```
prop.test(c(247,222),c(701,712),correct = F)
```

```
> prop.test(c(247,222),c(701,712),correct = F)

	2-sample test for equality of proportions without continuity correction

data:  c(247, 222) out of c(701, 712)
X-squared = 2.62, df = 1, p-value = 0.1055
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.008517967  0.089630022
sample estimates:
   prop 1    prop 2 
0.3523538 0.3117978 
```

By Hand

- Subtracting the sample estimates, we get

$$0.3523538 - 0.3117978 \approx 0.04$$

- We are 95% confident that the true difference of the two proportion is in the interval -0.01 to 0.09.
- The confidence interval contains zero, which makes it plausible that the true difference is zero.
- Looking at the p-value which is $0.11 > \alpha = 0.05$. 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 Gardasil
- $z = \sqrt{X^2} = \sqrt{2.62} = 1.62$

Chi-Square Test

What if we have more than two groups?

- **QUESTION**: Does the completion rate of the Gardasil vaccine vary by insurance type?
- To answer this, we have to compare the completion rates for the four groups of insurance types
  1. Hospital-Based
  2. Medical Assistance
  3. Military
  4. Private Player
- Since we have more than two groups, we can't just use a simple proportion test.
- We need to use the χ^2 test of association

`Completed` and `InsuranceType`

Use the `table()` command. Then, we use the `addmargins()`. Then, we can calculate the row propotion using the `prop.table()` command with the `margin` option set to 1.

Look at individuals on "medical assistance" and "hospital based" insurance.

```
> # Create a frequency table
> Insurance_Completion_Table <- table(gardasil$InsuranceType, gardasil$Completed)
> # View table
> Insurance_Completion_Table

                     no yes
  hospital based      45  39
  medical assistance 220  55
  military           209 122
  private payer      470 253
> # Add summary margines
> addmargins(Insurance_Completion_Table)

                       no  yes  Sum
  hospital based       45   39   84
  medical assistance  220   55  275
  military            209  122  331
  private payer       470  253  723
  Sum                 944  469 1413
> # Calculate the row proportions
> prop.table(Insurance_Completion_Table, margin = 1)

                            no       yes
  hospital based     0.5357143 0.4642857
  medical assistance 0.8000000 0.2000000
  military           0.6314199 0.3685801
  private payer      0.6500692 0.3499308
```

Using `chisq.test()`

- Medical assistance insurance have the lowest completion (20.0%) and those on hospital based insurance have the highest completion rate (46.4%).
- To run a χ^2 test, we can run the `chisq.test()` command. Both ways will give the same answer

```
chisq.test(gardasil$Completed,
gardasil$InsuranceType, correct = F)
chisq.test(Insurance_Completion_Table, correct =
F)
```

Using `chisq.test()`

```
> chisq.test(gardasil$Completed, gardasil$InsuranceType, correct = F)

        Pearson's Chi-squared test

data:  gardasil$Completed and gardasil$InsuranceType
X-squared = 31.283, df = 3, p-value = 7.411e-07

> chisq.test(Insurance_Completion_Table,correct = F)

        Pearson's Chi-squared test

data:  Insurance_Completion_Table
X-squared = 31.283, df = 3, p-value = 7.411e-07
```

Caution

An assumption of the χ^2 test is that all expected cell counts are at least 5. If we save the test as an object, we can test this.

```
> # Run and save chi square test for completion by insurance type
> Ins.Comp.test <- chisq.test(gardasil$Completed, gardasil$InsuranceType, correct = F)
> Ins.Comp.test$expected
                  gardasil$InsuranceType
gardasil$Completed hospital based medical assistance military private payer
               no         56.1189          183.72258 221.1352      483.0234
               yes        27.8811           91.27742 109.8648      239.9766
```

Fisher's Exact Test

Fisher's Exact Test

- An alternative to the χ^2 test when at least one expected cell count is less than 5. We can run the command `fisher.test()`.

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

	Fisher's Exact Test for Count Data

data:  gardasil$Completed and gardasil$InsuranceType
p-value = 3.432e-07
alternative hypothesis: two.sided

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

	Fisher's Exact Test for Count Data

data:  Insurance_Completion_Table
p-value = 3.432e-07
alternative hypothesis: two.sided
```

χ^2 Distribution

χ^2 Distribution

- Generally right-skewed
- To calculate the area under the curve, we use the `pchisq()` command.
- p-values based on a χ^2 test stat are given by the upper tail. BUT! by default, `pchisq()` gives the lower tail, so we need to take the complement.
- Example: when we examined `AgeGroup` and `Completion`, we got a χ^2 test stat of 2.62 on 1 degree of freedom

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

- We get 0.1055. No need to multiply this by 2 since we have a one tailed χ^2 p-value.