Module E: Advanced Inference

Slide Deck E7:

Testing Categorical Independence

The section in which we cover the Chi-square Test of Independence, which is used to test for independence between two categorical variables.

Today's Objectives

By the end of this slidedeck, you should

- understand the theory behind, and test hypotheses about:
  - determining if two categorical variables are independent
- better understand the p-value and how to test hypotheses

Test of Independence

Parametric Procedure: Chi-Square Test of Independence

- Null hypothesis: The two categorical variables are independent
- Graphics:
  - Matrix plot
    `mosaicplot( table(x,y) )`
  - Bar chart
    `barplot( table(x,y), beside=TRUE )`
- Requires: Expected number of successes is at least 10 in each table cell*
- `R` function: `chisq.test( table(x,y) )`

Note: This function *is* what Hawkes covers (with an adjustment). For `R` to give you Hawkes' results, you will need to use `correct=FALSE` in the function.

Framing Example

Example

I would like to test if there is a relationship between whether a person has blue eyes and whether that person is a MNS major.

To do this, I asked 100 people at Knox College and measure if they have blue eyes and if their major is in MNS. This is the contingency table:

	Blue	Not Blue
MNS Major	7	28
not MNS Major	13	52

Framing Example

Let's ask ourselves this question:

- What does it mean for two categorical variables to be independent?

It means:

- knowledge of one variable does not provide information about the other

In this specific example, it means that

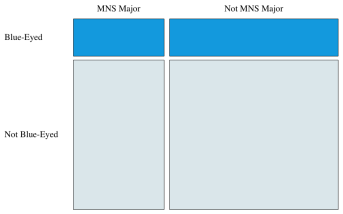
- the probability of a student being a MNS major does not depend on whether that student is blue-eyed or not

That is: The distribution of MNS majors is the same for blue-eyed as it is for non-blue-eyed students.

Let's check using the data...

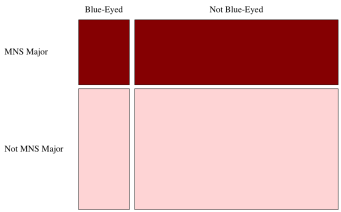
Framing Example

First, here is a mosaic plot for this data.



Framing Example

Here is a different mosaic plot (a different view) of this data.



Framing Example

The raw data:

	Blue	Not Blue
MNS Major	7	28
not MNS Major	13	52

What is the proportion of blue-eyed people who are MNS majors?

$$7/(7+13) = 35\%$$

What is the proportion of non-blue-eyed people who are MNS majors?

$$28/(28+52) = 35\%$$

It's a match! Therefore, knowledge of eye color gives no information about major.

Framing Example

The raw data:

	Blue	Not Blue
MNS Major	7	28
not MNS Major	13	52

What is the proportion of MNS majors who are blue-eyed?

$$7/(7+28) = 20\%$$

What is the proportion of non-MNS majors who are blue-eyed?

$$13/(13+52) = 20\%$$

It's a match! Therefore, knowledge of major gives no information about eye color.

Framing Example

In other words:

- Knowledge of whether the person is blue-eyed or not does not change the probability that person is a MNS major.
- Knowledge of a person's major does not change the probability that they are blue-eyed.

In other words: The two variables are independent.

Knowledge of one result does not affect the probability of the other.

The Test's Theory

Now that we've seen an example, let us generalize it. Here is some generic data:

	A_1	A_2	A_3
B_1	x_{11}	x_{12}	x_{13}
B_2	x_{21}	x_{22}	x_{23}

The Test's Theory

The data with column and row sums:

	A_1	A_2	A_3	Row Sum
B_1	x_{11}	x_{12}	x_{13}	r_1
B_2	x_{21}	x_{22}	x_{23}	r_2
Column Sum	c_1	c_2	c_3	n

The Test's Theory

If the two variables are perfectly independent, the **expected values** would be:

	A_1	A_2	A_3	Row Sum
B_1	$\mu_{11} = n \cdot \frac{r_1}{n} \cdot \frac{c_1}{n}$	$\mu_{12} = n \cdot \frac{r_1}{n} \cdot \frac{c_2}{n}$	$\mu_{13} = n \cdot \frac{r_1}{n} \cdot \frac{c_3}{n}$	r_1
B_2	$\mu_{21} = n \cdot \frac{r_2}{n} \cdot \frac{c_1}{n}$	$\mu_{22} = n \cdot \frac{r_2}{n} \cdot \frac{c_2}{n}$	$\mu_{23} = n \cdot \frac{r_2}{n} \cdot \frac{c_3}{n}$	r_2
Column Sum	c_1	c_2	c_3	n

The Test's Theory

The previous slide stated

If the two variables are perfectly independent, the expected values would be...

This comes from one definition of independent events:

Events A and B are independent if $\mathbb{P}[\ A \text{ and } B\ ] = \mathbb{P}[\ A\ ] \times \mathbb{P}[\ B\ ]$

Thus, under the null hypothesis that the two variables *are* independent, the probability of a person being in the A_1B_1 cell is the probability of being A_1 times the probability of being B_1.

And, the expected *number* of people in that cell is just n times that probability:

$$\mu_{11} = n\ \mathbb{P}[\ A_1\ ]\ \mathbb{P}[\ B_1\ ] = n\ \frac{c_1}{n}\ \frac{r_1}{n}$$

The Test's Theory

However, since the sample is random, two independent variables cannot always give such perfect results.

- Thus, we need to create a test statistic to determine if the differences are likely due to random chance or something systematic.

As before, that test statistic is

$$\text{X2} = \sum \frac{(x_i - \mu_i)^2}{\mu_i}$$

As before in the goodness-of-fit test, we are summing up over all of the cells.

The test statistic has a Chi-square distribution with

$$df = (r-1)(c-1)$$

degrees of freedom.

Framing Example, Continued

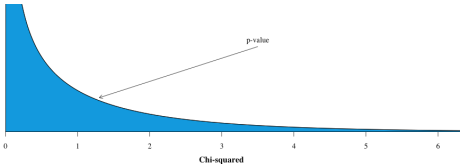
Continuing the framing example:

$$\begin{aligned}
\text{X2} &= \sum \frac{(x_i - \mu_i)^2}{\mu_i} \\
&= \frac{(7 - 20 \times 35/100)^2}{20 \times 35/100} + \frac{(28 - 80 \times 35/100)^2}{80 \times 35/100} + \frac{(13 - 20 \times 65/100)^2}{20 \times 65/100} + \frac{(52 - 65 \times 80/100)^2}{65 \times 80/100} \\
&= \frac{(7-7)^2}{7} + \frac{(28-28)^2}{28} + \frac{(13-13)^2}{13} + \frac{(52-52)^2}{52} \\
&= \frac{0}{7} + \frac{0}{28} + \frac{0}{13} + \frac{0}{52} \\
&= 0
\end{aligned}$$

This test statistic follows a Chi-square distribution with $(2-1)(2-1) = 1$ degree of freedom.

Framing Example, Continued

This illustrates the p-value. Since the value of the test statistic is 0, the entire graphic is shaded, meaning the p-value is 1.00.



Example 1: Gendered Hat Fashion

Example

I would like to determine if the proportion of males who wear hats is the same as the proportion of females who do. To test this, I sample 100 males and 100 females. Ten males and 16 females were wearing hats.

Notes:

- we have already seen this example in the context of comparing two proportions
- it is appropriate to investigate this as a question of whether the two variables (gender and hat-wearing) are independent
- none of those sampled identified as non-binary

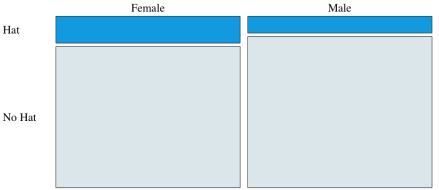
Example 1: Gendered Hat Fashion

The table of observed values

	Hat	no Hat	Row Sum
Female	16	84	100
Male	10	90	100
Column Sum	26	174	200

Example 1: Gendered Hat Fashion

This is a mosaic plot of the data. Again, be able to interpret the graphic.



Example 1: Gendered Hat Fashion

The table of expected values is

	Hat	no Hat	Row Sum
Female	$200 \cdot \frac{100}{200} \cdot \frac{26}{200}$	$200 \cdot \frac{100}{200} \cdot \frac{174}{200}$	100
Male	$200 \cdot \frac{100}{200} \cdot \frac{26}{200}$	$200 \cdot \frac{100}{200} \cdot \frac{174}{200}$	100
Column Sum	26	174	200

Example 1: Gendered Hat Fashion

The table of expected values is

	Hat	no Hat	Row Sum
Female	13	87	100
Male	13	87	100
Column Sum	26	174	200

Example 1: Gendered Hat Fashion

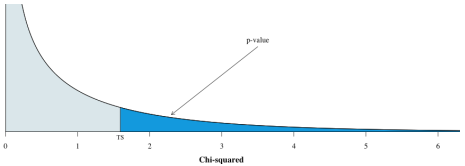
The test statistic value is

$$\begin{aligned}
\text{X2} &= \sum_i \frac{(x_i - \mu_i)^2}{\mu_i} \\
&= \frac{(16-13)^2}{13} + \frac{(84-87)^2}{87} + \frac{(10-13)^2}{13} + \frac{(90-87)^2}{87} \\
&= \frac{9}{13} + \frac{9}{87} + \frac{9}{13} + \frac{9}{87} \\
&= 0.6923 + 0.1034 + 0.6923 + 0.1034 \\
&= 1.5915
\end{aligned}$$

This test statistic follows a Chi-square distribution with $(2-1)(2-1) = 1$ degree of freedom.

Example 1: Gendered Hat Fashion

This illustrates the p-value for a Chi-squared distribution of 1 degree of freedom and a test statistic of 1.5915.



Example 1: Gendered Hat Fashion

The p-value is

$$\begin{aligned} \text{p-value} &= \mathbb{P}[\ X2 \geq 1.5915\ ] \\ &= \texttt{1-pchisq(1.5915, df=1)} \\ &= 0.2071 \end{aligned}$$

Brief Conclusion:
Because the p-value of 0.2071 is greater than our usual alpha of 0.05, we cannot reject the null hypothesis. There is no significant evidence that the hat-wearing rate differs between males and females.

```
obs = matrix( c(16,10,84,90), ncol=2)
chisq.test(obs, correct=FALSE)
```


Example 1: Gendered Hat Fashion

The R code and output for the "better" test is

```
obs = matrix( c(16,10,84,90), ncol=2)
chisq.test(obs)
```

Output:

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  obs
X-squared = 1.1052, df = 1, p-value = 0.2931
```

Note: The difference between the two tests is minor when the sample size is large.

Example 2: Grade Fairness by Major Type

Example

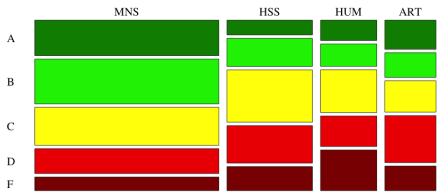
I would like to determine if different major types have a different grade distribution in my STAT 200 sections.

Observations from past courses:

Major Type	A	B	C	D	F
MNS	57	72	61	40	22
HSS	11	21	39	28	18
HUM	10	11	21	15	20
ART	13	11	14	21	11

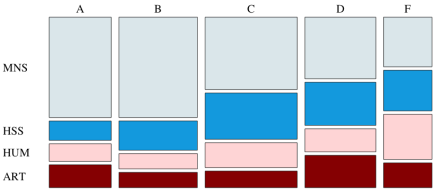
Example 2: Grade Fairness by Major Type

A mosaic plot for these data



Example 2: Grade Fairness by Major Type

A different mosaic plot for these data



Example 2: Grade Fairness by Major Type

The R code for the Hawkes test is

```
grades = matrix(
  c(57,72,61,40,22, # MNS
    11,21,39,28,18, # HSS
    10,11,21,15,20, # HUM
    13,11,14,21,11  # ART
    ), ncol=5, byrow=TRUE )

rownames(grades) = c("MNS","HSS","HUM","ART")
colnames(grades) = c("A","B","C","D","F")

chisq.test(grades)
```

Example 2: Grade Fairness by Major Type

The resulting output:

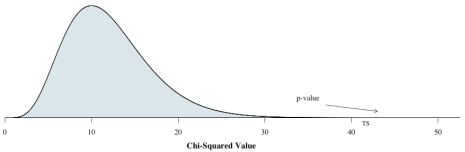
```
  Pearson's Chi-squared test

data:  grades
X-squared = 41.672, df = 12, p-value = 3.781e-05
```

Brief Conclusion:
Because the p-value of 3.781×10^{-5} is less than our usual alpha of 0.05, we should reject the null hypothesis. There is significant evidence that the the grade distribution differs among the major types.

Example 2: Grade Fairness by Major Type

This illustrates the p-value for a Chi-squared distribution of 12 degrees of freedom and a test statistic of 41.672.



Example 3: The Democratic Peace Thesis

Example

The Democratic Peace Thesis states that democratic countries are more peaceful than non-democratic countries. It was assumed to be common sense until the early 1970s, when researchers decided to test it and discover that it was not true. This has led to 50+ years of researchers determining under which conditions it *is* true.

In this example, let us define "peaceful" as a country *not* being the first users of force in a militarized interstate dispute (MID). The data set includes all MIDs from the 1960s until the mid-2000s.

Let us determine if there is a relationship between government type and whether or not the country was the first user of force in a MID.

Example 3: The Democratic Peace Thesis

Here is a contingency table of the data

Government Type	First User	Not First User
Democracy	128	590
Anocracy	77	241
Autocracy	220	368

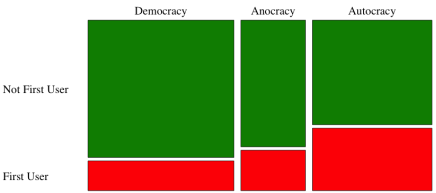
I obtained this table using the following code:

```
dt = read.csv("https://rfs.kvasaheim.com/data/fuf.csv")
attach(dt)

dataTable = table(govtType, fuffer)
colnames(dataTable) = c("Not First User", "First User")
```

Example 3: The Democratic Peace Thesis

A mosaic plot for these data



Example 3: The Democratic Peace Thesis

The R code and output for the test is

```
chisq.test(dataTable)
```

The resulting output:

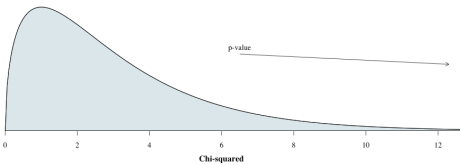
```
Pearson's Chi-squared test

data:  dataTable
X-squared = 64.976, df = 2, p-value = 7.773e-15
```

Brief Conclusion:
Because the p-value of 7.77×10^{-15} is much smaller than our usual alpha of 0.05, we should reject the null hypothesis. There is is significant evidence of a relationship between government type and whether that country was the first user of force in a MID.

Example 3: The Democratic Peace Thesis

This illustrates the p-value for a Chi-squared distribution of 3 degrees of freedom and a test statistic of 64.976.



Today's Objectives

Now that we have concluded this lecture, you should be able to

1. understand the theory behind, and test hypotheses about:
   - determining if two categorical variables are independent
2. better understand the p-value and how to test hypotheses

Since we used R to perform the calculations, we were better able to focus on the interpretation than on the tedious calculations.

As always: Please do not forget to be familiar with the `allProcedures` document that lists all of the statistical procedures we will use in R.

Today's R Functions

Here is what we used the following three R functions. The first performs the test, while the second illustrates the data. Both of these require the data to be in the form of a matrix (the third function).

- `chisq.test(m)`
  This performs the Chi-square test of Independence (make sure `m` is a matrix)
- `mosaicplot(m)`
  This produces a basic mosaic plot (again, make sure `m` is a matrix)
- `matrix(x, ncol)`
  This creates a matrix of the data (entered by column).

Supplemental Activities

The following activity is currently available from the STAT 200 website to give you some practice in performing hypothesis tests concerning the Chi-square Test of Independence.

- SCA 41

Source: `https://www.kvasaheim.com/courses/stat200/sca/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of understanding how to test if two categorical variables are independent:

- Hawkes Learning: Section 10.7
- Intro to Modern Statistics: Chapter 18
- R for Starters: None

- Wikipedia: Hypothesis testing
  Chi-squared test

Please do not forget to use the `allProcedures` document that lists all of the procedures we will use in R.