Chi-Squared Analysis

Dr. J. Kyle Roberts

Southern Methodist University
Simmons School of Education and Human Development
Department of Teaching and Learning

Background to Chi-Squared

- Frequently data are presented to us as counts.
- The number of people with a certain characteristic.
- The number of students who do not graduate.
- The number of patients who die.
- In the Chi-squared (χ^2) analysis, we consider our data in a contingency table and compare the "observed" frequencies against the "expected" frequencies.

Heuristic Data for χ^2 Analysis

- Suppose that we have the following dataset where we are sampling people and collect two pieces of information.
- Whether their eyes are "blue" or "brown"
- AND whether their hair is "fair" or "dark"

	Blue eyes	Brown eyes
Fair hair	38	11
Dark hair	14	51

- Using this data, we can produce the row and column totals:

	Blue eyes	Brown eyes	Row Totals
Fair hair	38	11	49
Dark hair	14	51	65
Column totals	52	62	114

Computing the "Expected" Frequencies

- We will first compute the expected frequency for having "fair" hair and "blue" eyes.
- Since it is assumed that having "fair" hair and "blue" eyes are independent factors, then we compute the probability of having both as the product of the probability of having each.
- For example, the probability of having blue eyes is $52/114 = 0.456$, and the probability of having fair hair is $49/114 = 0.430$.
- Then it follows that the "expected" probability of having both traits is $(52/114) * (49/114) * 114 = 22.35$.
- Solving for all probabilities, we obtain:

	Blue eyes	Brown eyes	Row Totals
Fair hair	22.35	26.65	49
Dark hair	29.65	35.35	65
Column totals	52	62	114

Computing the Pearson χ^2

- The test statistic is:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- where O is the observed frequency and E is the expected frequency

	O	E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Fair/blue	38	22.35	244.92	10.96
Fair/brown	11	26.65	244.92	9.19
Dark/blue	14	29.65	244.92	8.26
Dark/brown	51	35.35	244.92	6.93

- Thus we would calculate χ^2 as $10.96 + 9.19 + 8.26 + 6.93 = 35.34.$

df for χ^2

- For a given contingency table, the df are calculated as a product of the number of columns and number of rows.
- Thus for our data:

$$\begin{aligned} df &= (r-1)(c-1) \\ &= (2-1)(2-1) \\ &= 1 \end{aligned}$$

- Based on a $df = 1$ we can compute a critical value for χ^2 at $\alpha = 0.05$ by:

```
> qchisq(0.95, 1)
[1] 3.841459
```

Determining Whether or Not to Reject H_0

- Since our observed $\chi^2_{calc} = 35.33 > \chi^2_{crit} = 3.84$ we reject the H_0 that eye color and hair color are independent.
- Since we also know the distribution of χ^2, we can compute the probability of the null being "true" in R.

```
> haireye <- data.frame(eyes = rep(c("blue", "brown"),
+     c(52, 62)), hair = rep(c("fair", "dark", "fair",
+     "dark"), c(38, 14, 11, 51)))
> table(haireye)
```

```
       hair
eyes     dark fair
  blue     14   38
  brown    51   11
```

χ^2 in R

```
> chisq.test(table(haireye), correct = F)

Pearson's Chi-squared test
data:  table(haireye)
X-squared = 35.3338, df = 1, p-value = 2.778e-09

> chisq.test(table(haireye), correct = F)$expected

       hair
eyes        dark     fair
  blue  29.64912 22.35088
  brown 35.35088 26.64912

> table(haireye)

       hair
eyes    dark fair
  blue    14   38
  brown   51   11
```

This means that there is a positive relationship between “blue” eyes and “fair” hair.

Standardized Residuals in χ^2

- We can determine which of the categories are major contributors to the statistically significant χ^2 by computing the standardized residual.

$$R = \frac{O - E}{\sqrt{E}}$$

	O	E	$(O-E)$	R
Fair/blue	38	22.35	15.65	3.31
Fair/brown	11	26.65	-15.65	-3.03
Dark/blue	14	29.65	-15.65	-2.87
Dark/brown	51	35.35	15.65	2.63

```
> chisq.test(table(haireye), correct = F)$resid

       hair
eyes         dark      fair
  blue  -2.873982  3.310112
  brown  2.632024 -3.031437
```

Adding Another Contingency

- Suppose that `haireye` has another contingency such as `gender`
- We may want to see if there are any additional differences among the expected frequencies of hair and eye color among males and females

```
> set.seed(12346)
> haireye$gender <- sample(0:1, 114, replace = T)
> table(haireye)

, , gender = 0
       hair
eyes    dark fair
  blue    10   20
  brown   23    4

, , gender = 1

       hair
eyes    dark fair
  blue     4   18
  brown   28    7
```

χ^2 with Gender Contingency

```
> (mnew <- chisq.test(table(haireye), correct = F))

Chi-squared test for given probabilities
data:  table(haireye)
X-squared = 41.6491, df = 7, p-value = 6.075e-07

> mnew$resid

, , gender = 0
       hair
eyes         dark      fair
  blue  -1.1258525  1.5232122
  brown  2.3179316 -2.7152913

, , gender = 1

       hair
eyes         dark      fair
  blue  -2.7152913  0.9933993
  brown  3.6424640 -1.9205719
```

χ^2 without a Dataset

- You can also run a χ^2 if you just have the frequencies and do not have the actual dataset. For example:

	Males	Females
Dropped Out	32	24
Finished HS	265	199
Went to College	391	287

```
> observed <- matrix(c(32, 24, 265, 199, 391, 287),
+      nrow = 3, byrow = T)
> chisq.test(observed, correct = F)

Pearson's Chi-squared test
data:  observed
X-squared = 0.037, df = 2, p-value = 0.9817

> cbind(observed, chisq.test(observed)$resid)

     [,1] [,2]        [,3]        [,4]
[1,]   32   24 -0.02826081  0.03282417
[2,]  265  199 -0.09009990  0.10464860
[3,]  391  287  0.08265842 -0.09600552
```