Module 8

Categorical Associations

James Baglin

Last updated: 13 July, 2020

Overview

##Summary

Statistical investigations often explore the distribution of categorical variables or explore the associations between categorical variables. This module introduces hypothesis testing for categorical variables based on the Chi-square distribution.

Learning Objectives

The learning objectives associated with this module are:

- Determine when a Chi-square goodness of fit test or Chi-square test of association may be applied to categorical data.
- Apply and interpret the Chi-square goodness of fit test.

- Apply and interpret the Chi-square test of association.
- Use technology to compute a Chi-square goodness of fit test and Chi-square test of association.
- Interpret a Chi-square goodness of fit test and Chi-square test of association.

Video

Chi-square Goodness of Fit Test

Categorical data are all around us. Gender, car make, food type, blood type, and hair colour are a few examples. Continuous data are also often converted to categorical or ordinal bands (e.g. age bands, weight bands, or income bands) to simplify analysis and presentation. In this module, we will introduce the Chi-square goodness of fit test and Chi-square test of association as an introduction to Hypothesis testing with categorical variables.

M&Ms Data

Milk chocolate M&M's are supposedly manufactured to contain 24% blue, 13% brown, 16% green, 20% orange, 13% red, and 14% yellow M&M's. Suppose someone, perhaps with a little too much time on their hands, purchases 48 packs of M&Ms (https://joshmadison.com/2007/12/02/mms-color-distribution-analysis/) and collates the following data:

Colour	Blue	Brown	Green	Orange	Red	Yellow	Total
Observed	481	371	483	544	372	369	2620
Proportion	0.184	0.142	0.184	0.208	0.142	0.141	
Expected	629	341	419	524	341	367	

Colour	Blue	Brown	Green	Orange	Red	Yellow	Total
π	0.24	0.13	0.16	0.20	0.13	0.14	

The observed row reports the count of each colour tallied from the 48 M&M packs. The proportion row reports the observed proportions. For example, 481 of the sampled M&Ms were blue. The proportion of observed blue M&Ms was therefore, 481/2620 = 0.184. The expected row reports the number of M&Ms that we would expect to be a particular colour assuming the manufacturer's claims are true, which are recorded in the Population Proportion row. For example, if M&Ms are manufactured with 24% blue, we would expect 2620*.24 = 629 M&Ms to be blue in our sample. We notice that blue M&Ms are under-represented (observed = 481), which means the other colours are all slightly over-represented. Does this mean the manufacturer is wrong?

We need to be careful before jumping to this conclusion as we have used a sample. We need to perform hypothesis testing on the data to determine if this under-representation of blue M&Ms can be considered statistically significant. We can perform a Chi-square goodness of fit test to do just that.

The **Chi-square goodness of fit test** is a hypothesis test in which we test if the count of our categorical observations "fit" the model of expected outcomes. In this example, the expected outcomes are based on the manufacturer's claimed proportions.

The Chi-square goodness of fit test has one major assumption. **The minimum expected value for each cell must be at least 5**. According to the table above, this assumption was clearly met.

The Chi-square goodness of fit test is based on the calculation of the Chi-square statistic:

$$\chi^2 = \sum \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$$

where Obs are the observed values and Exp are the expected values based on the population probability distribution.

Let's calculate the Chi-square statistic, χ^2:

$$\chi^2 = \frac{(481-629)^2}{629} + \frac{(371-341)^2}{341} + \frac{(483-419)^2}{419} + \frac{(544-524)^2}{524} + \frac{(372-341)^2}{341} + \frac{(369-367)^2}{367} = \frac{-147.8^2}{629} + \frac{30.4^2}{341} + \frac{63.8^2}{419} + \frac{20^2}{524} + \frac{31.4^2}{341} + \frac{2.2^2}{367} = \frac{21844.8}{629} + \frac{924.2}{341} + \frac{4070.4}{419} + \frac{400}{524} + \frac{986}{341} + \frac{4.8}{367} = 34.74 + 2.71 + 9.71 + 0.76 + 2.89 + 0.01 = 50.82$$

As you can see, the χ^2 reflects the degree of discrepancy between the observed and expected values. If the observed values were exactly the expected values, the χ^2 statistic would be 0.

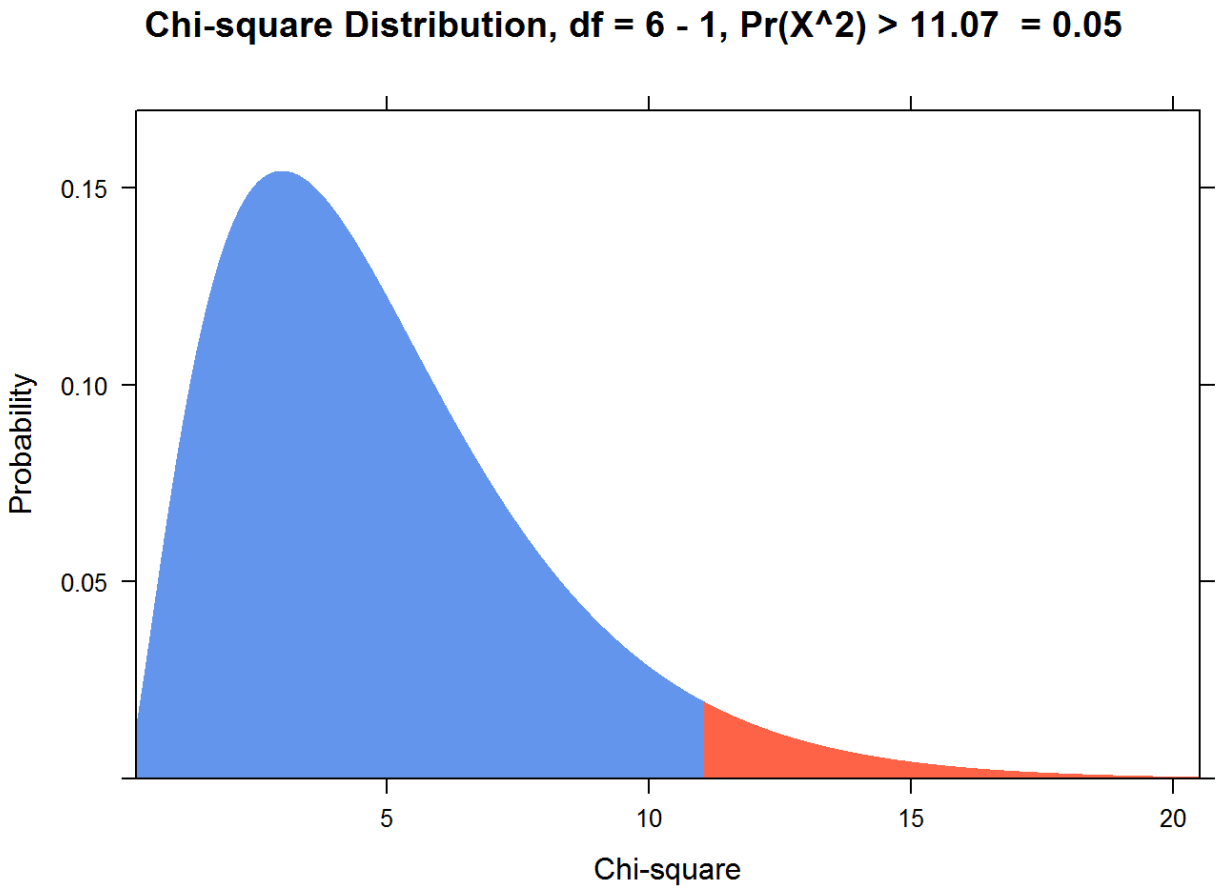
We can now use this Chi-square statistic to test the following hypotheses at the standard 0.05 significance level:

H_0: The population distribution of M&M colours are 24% blue, 13% brown, 16% green, 20% orange, 13% red, and 14% yellow.
H_A: The population distribution of M&M colours are NOT 24% blue, 13% brown, 16% green, 20% orange, 13% red, and 14% yellow.

We can use a χ^2 critical value or p-value to decide whether to reject or fail to reject H_0.

The critical value is found by looking up the value on a Chi-square distribution, with df = Number of categories - 1 = 6 - 1 = 5, that satisfies $Pr(\chi^2 > x) = .05$. Use the following app to find this value. Select Chi-square as your Distribution, $df = 5$, Defined Shaded Area By: Probability, Area to shade: Right tail and enter .95 as the probability.

You should be able to confirm $Pr(\chi^2 > 11.07) = .05$, which is depicted in the following plot.



We can also use the R function `qchisq()`.

```
qchisq(p = 0.05, df = 5, lower.tail = FALSE)
```

```
## [1] 11.0705
```

This tells us that χ^2 values above 11.07 have less than 0.05 probability of occurring assuming H_0 is true. As $\chi^2 = 50.83$, which is greater than the critical value 11.07, the decision was to reject H_0.

An easier way to test H_0 would be to calculate the p-value. The p-value can be calculated as $Pr(\text{X2} > 50.83)$. Using R:

```
pchisq(q = 50.83, df = 5,lower.tail = FALSE)
```

```
## [1] 9.370786e-10
```

This p-value is really small. We should round this to $p < .001$. This p-value is less than the standard significance level of 0.05. Therefore, we reject H_0. Both the critical value and p-value approach come to the same conclusion. The p-value approach is the more

informative. The results of the Chi-square goodness of fit test found a statistically significant departure from the proportion of M&M colours claimed to be produced by the manufacturer.

Chi-square Goodness of Fit Test using R

Download the `m&m.csv` (data/m&ms.csv) dataset or create a `data.frame()` using the following code:

```
Col<-c("Blue","Brown","Green","Orange","Red","Yellow")
obs<-c(481,371,483,544,372,369)
mms<-data.frame(Colour=rep(Col,obs))
```

You can summarise the data using the `table()` and `prop.table()` functions:

```
table(mms$Colour)
```

```
##
##   Blue  Brown  Green Orange    Red Yellow
##    481    371    483    544    372    369
```

```
table(mms$Colour) %>% prop.table()
```

```
##
##      Blue     Brown     Green    Orange       Red    Yellow
## 0.1835878 0.1416031 0.1843511 0.2076336 0.1419847 0.1408397
```

Next, define your population proportions (ensure these sum to 1!)

```
pop_prop<-c(0.24,0.13,0.16,0.20,0.13,0.14)
```

Now run the Chi-square goodness of fit test using the previous `table()` and the `chisq.test()` function:

```
chi1<-chisq.test(table(mms$Colour), p = pop_prop)
chi1
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(mms$Colour)
## X-squared = 50.835, df = 5, p-value = 9.348e-10
```

```
# Observed
chi1$observed
```

```
##
##   Blue  Brown  Green Orange    Red Yellow
##    481    371    483    544    372    369
```

```
# Expected
chi1$expected
```

```
##   Blue  Brown  Green Orange    Red Yellow
##  628.8  340.6  419.2  524.0  340.6  366.8
```

Note that the p-value and χ^2 value are slightly different to the previously worked example due to decimal rounding.

Example Write-up

A Chi-square goodness of fit test was used to determine whether the distribution of M&M colours followed the manufacturer's purported distribution of 24% blue, 13% brown, 16% green, 20% orange, 13% red, and 14% yellow M&M's. The test was statistically significant, χ^2 = 50.83, $df = 5, p < .001$. This suggests that the distribution of M&M colours do not follow the distribution proposed by the manufacturer.

Limitations

While this investigation considered real data, it's important to note that there was an issue with the original sample. The investigator used 48 packs of M&Ms purchased from the same store. Could this be considered a random sample of all M&Ms? Probably not. It's possible that the packets purchased were biased. Perhaps on that particular day or time, the machines packaging the M&Ms might have been a little lax with the blue M&Ms. The investigator would need to follow-up with a proper random sample and repeat the Chi-square goodness of fit test before being confident in their conclusion.

Chi-square Test of Association

Breast Cancer Data

Consider the following real world scenario. An investigator gathers a random sample of 3,220 mothers with a diagnosis of breast cancer (cases) and 10,245 mothers without a history of breast cancer (controls). All mothers are then asked to report their age at the birth of their first child. The investigator presents this data in a 2 (rows) x 5 (columns) contingency table:

Mother's Age at First Birth

Case-control Status		<20	20-24	25-29	30-34	>=35	Total
Case	Count	320	1206	1011	463	220	3220
(Breast cancer)	% Within Group	9.9%	37.5%	31.4%	14.4%	6.8%	100.0%
	% Within Age	18.4%	21.4%	25.9%	29.8%	35.1%	23.9%
Control	Count	1422	4432	2893	1092	406	10245
(No breast cancer)	% Within Group	13.9%	43.3%	28.2%	10.7%	4.0%	100.0%
	% Within Age	81.6%	78.6%	74.1%	70.2%	64.9%	76.1%
Total	Count	1742	5638	3904	1555	626	13465
	% Within Group	12.9%	41.9%	29.0%	11.5%	4.6%	100.0%
	% Within Age	100%	100%	100%	100%	100%	100%

The table presents the raw counts, % within group and % within age categories. The raw counts tell us how many mothers were observed. For example, there were 2,893 controls in the 25-29 age category and 463 cases aged 30 - 34 at the time of their first child. The % within group tells us what percentage of mothers within either the cases or controls that belonged to a certain age category. For example, 6.8% of cases were aged >= 35 years vs. 4.0% of controls. The % within age reports the percentage of cases or controls within the age category. For example, 70.2% of mothers in the 30-34 age category were controls. We can present the % within groups using a clustered barchart using R. The data from the table above are available in the Breast Cancer.csv (data/Breast%20Cancer.csv) dataset. First, import the data and label the variables.

```
Breast.Cancer <- read.csv("data/Breast Cancer.csv")

Breast.Cancer$Group <- factor(Breast.Cancer$Group, levels=c(1,2),
                              labels = c("Case","Control"))

Breast.Cancer$Age_Cat <- factor(Breast.Cancer$Age_Cat, levels = c(1,2,3,4
  ,5),
                              labels = c("< 20","20 - 24","25 - 29",
  "30 - 34", ">= 35"),
                              ordered=TRUE)
```

Now let's table the data.

```
table(Breast.Cancer$Age_Cat, Breast.Cancer$Group)
```

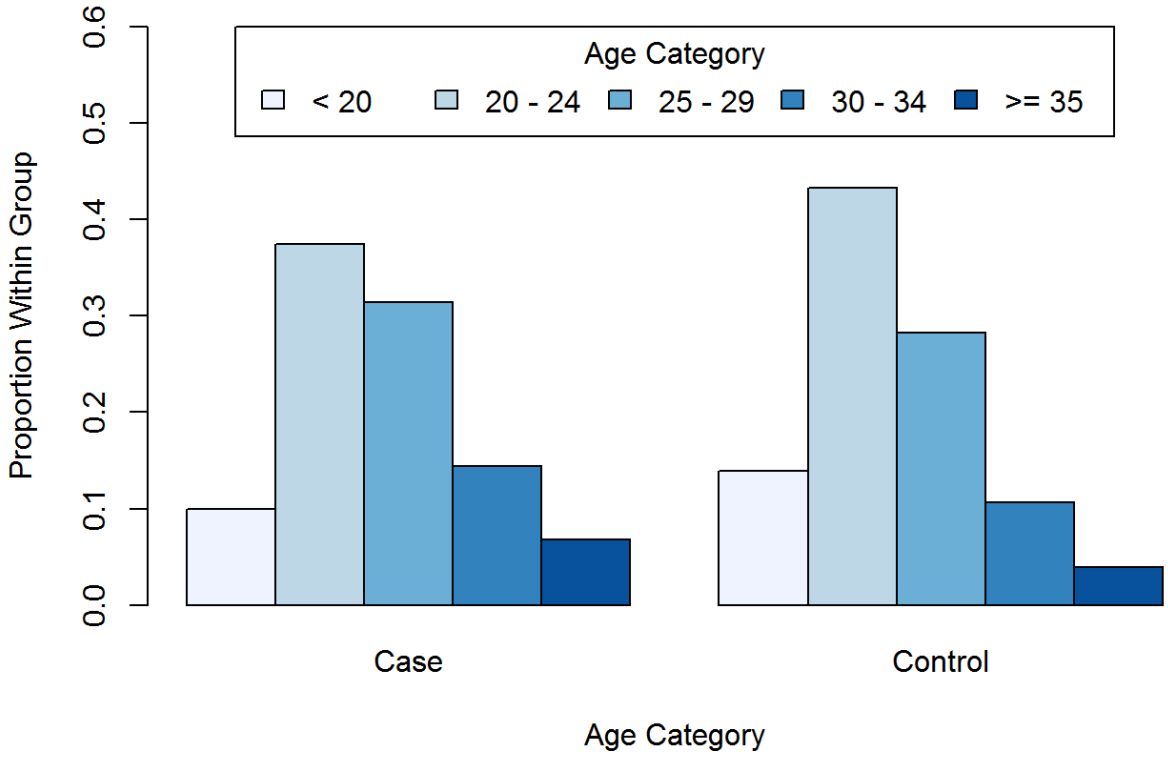
```
##
##           Case Control
##   < 20     320    1422
##   20 - 24 1206    4432
##   25 - 29 1011    2893
##   30 - 34  463    1092
##   >= 35    220     406
```

```
table(Breast.Cancer$Age_Cat, Breast.Cancer$Group) %>% prop.table(margin =
  2)
```

```
##
##                 Case    Control
##   < 20    0.09937888 0.13879941
##   20 - 24 0.37453416 0.43260127
##   25 - 29 0.31397516 0.28238165
##   30 - 34 0.14378882 0.10658858
##   >= 35   0.06832298 0.03962909
```

Now for a nice visual summary to help explain the potential association.

```
library(RColorBrewer)
table <- table(Breast.Cancer$Age_Cat, Breast.Cancer$Group) %>% prop.table
  (margin = 2)
barplot(table,ylab="Proportion Within Group",
          ylim=c(0,.6),legend=rownames(table),beside=TRUE,
          args.legend=c(x = "top",horiz=TRUE,title="Age Category"),
          xlab="Age Category", col = brewer.pal(5, name = "Blues"))
```



If there was no association between age at first birth and breast cancer, the height of the bars (i.e. proportions) of cases and controls within each of the age bands would be the same. In the barchart above, this does not seem to be the case. The cases (i.e. mothers with breast cancer) are less likely to have their children under the age of 25 when compared to controls. This is an example of a categorical association. In other words, the probability of being a case "depends" on the age of the mother when they first give birth. What we need to determine with a Chi-square test of association is whether this relationship is statistically significant or whether it reflects natural sampling variability assuming breast cancer and age are independent (i.e. H_0). The statistical hypotheses for this Chi-square test of association can be written as follows:

H_0: There is no association in the population between the categorical variables (independence)
H_A: There is an association in the population between the categorical variables (dependence)

To test H_0, the Chi-square statistic is calculated as:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed count in the i^{th} row of the j^{th} column and E_{ij} is the expected count assuming no association. E_{ij} is calculated as...

$$E_{ij} = n\left(\frac{r_i}{n}\right)\left(\frac{c_j}{n}\right)$$

where r_i refers to the total count of the i^{th} row and c_j is the total count of the j^{th} column. For example...

$$E_{1,1} = 13465\left(\frac{3220}{13465}\right)\left(\frac{1742}{13465}\right) = 416.6$$

and...

$$E_{1,2} = 13465\left(\frac{3220}{13465}\right)\left(\frac{5638}{13465}\right) = 1348.3$$

Fortunately, we can get R to compute and report the rest using the `chisq.test()` function:

Chi-square Test of Association using R

```
chi2 <- chisq.test(table(Breast.Cancer$Group,Breast.Cancer$Age_Cat))
chi2
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(Breast.Cancer$Group, Breast.Cancer$Age_Cat)
## X-squared = 130.34, df = 4, p-value < 2.2e-16
```

```
# Observed
chi2$observed
```

```
##
##           < 20 20 - 24 25 - 29 30 - 34 >= 35
##   Case     320    1206    1011     463   220
##   Control 1422    4432    2893    1092   406
```

```
# Expected
chi2$expected
```

```
##
##                < 20  20 - 24   25 - 29   30 - 34    >= 35
##   Case     416.5793 1348.263  933.5967  371.8604 149.7007
##   Control 1325.4207 4289.737 2970.4033 1183.1396 476.2993
```

The Chi-square test of association assumes that no more than 25% of the cells have expected counts below 5. This assumption is met in the above table as all expected counts (expected) are above 5. Using the complete list of expected counts, we can calculate the χ^2 statistic:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(320 - 416.6)^2}{416.6} + \frac{(1206 - 1348.3)^2}{1348.3} + \frac{(1101 - 933.6)^2}{933.6} + \frac{(463 - 371.9)^2}{371.9} + \frac{(220 - 149.7)^2}{149.7} + \frac{(1422 - 1325.4)^2}{1325.4} + \frac{(4432 - 4289.7)^2}{4289.7} + \frac{(2893 - 2970.4)^2}{2970.4} + \frac{(1092 - 1183.1)^2}{1183.1} + \frac{(406 - 476.3)^2}{476.3} = 130.338$$

This is better left to R. The Chi-square statistic represents the degree of discrepancy between the observed and expected counts under H_0. This statistic follows a χ^2 distribution with

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

therefore, $df = (2 - 1)(5 - 1) = 4$.

Using a critical value approach to test H_0, we need to compare the observed $\chi^2 = 130.338$ to a critical value, χ^2_{crit}. The Chi-square distribution is a positive distribution, therefore, we perform a one-tailed hypothesis test. The one-tailed critical value, χ^2_{crit}, associated with $\alpha = 0.05$ and $df = 4$ can be calculated using the following R formula:

```
qchisq(p = .95,df = 4)
```

```
## [1] 9.487729
```

The critical value was found to be 9.49. We reject H_0 when $\chi^2 > \chi^2_{crit}$. As 130.338 > 9.49, H_0 was rejected. As R does not report the critical value, the easiest method to test H_0 is to use the p-value. To calculate the p-value, we need to find the probability of observing the sample χ^2, or one more extreme, assuming there is no association in the population. We can use the `pchisq()` function:

```
pchisq(q = 130.34,df = 4,lower.tail = FALSE)
```

```
## [1] 3.293743e-27
```

We find $p < .001$. The output from the `chisq.test()` function above reports these values by default. However, you will notice that the p-value was reported to be p-value < 2.2e-16. This is a limitation of the printout. To confirm the p-value computed above, run the following code:

```
chi<-chisq.test(table(Breast.Cancer$Group,Breast.Cancer$Age_Cat))
chi$p.value
```

```
## [1] 3.29696e-27
```

This code saves the `chisq.test()` as an object named " `chi` " and then reports the `p.value` saved with this object. This is a similar p-value derived using `pchisq()`.

As this p-value was less than the 0.05 level of significance, H_0 was rejected. There was a statistically significant association between breast cancer and the age at which a mother first gives birth.

Example Write-up

A Chi-square test of association was used to test for a statistically significant association between breast cancer status and the age of a mother at first birth. The results of the test found a statistically significant association, $\chi^2 = 130.34, p < .001$. The results of this study suggest that women with breast cancer were more likely to give birth to their first child in older age categories when compared to controls.

References