Chi-Squared Analysis

Lynn Waterhouse

March 23, 2017

Chi-Squared Analysis

There are a variety of chi-square tests. -Chi-square goodness of fit -Chi-square test for independence

Chi-Square Goodness of fit

The chi-square test (Snedecor and Cochran, 1989) is used to test if a sample of data came from a population with a specific distribution.

This is a chi-square distribution with (k - c) degrees of freedom where k is the number of non-empty cells and c = the number of estimated parameters (including location and scale parameters and shape parameters) for the distribution + 1.

Assumptions for Chi-Square Goodness of Fit

The chi-square goodness-of-fit test is applied to binned data (i.e., data put into classes). This is actually not a restriction since for non-binned data you can simply calculate a histogram or frequency table before generating the chi-square test. However, the value of the chi-square test statistic are dependent on how the data is binned. Another disadvantage of the chi-square test is that it requires a sufficient sample size in order for the chi-square approximation to be valid.

For the chi-square approximation to be valid, the expected frequency should be at least 5. This test is not valid for small samples, and if some of the counts are less than five, you may need to combine some bins in the tails.

Chi-Square Test for Independence

The test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

Degrees of freedom. The degrees of freedom (DF) is equal to: DF = (r - 1) * (c - 1)

where r is the number of levels for one catagorical variable, and c is the number of levels for the other categorical variable.

When to Use Chi-Square Test for Independence:

-The sampling method is simple random sampling. -The variables under study are each categorical. -If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5.

Using R to calculate

use chisq.test

```
?chisq.test
```

```
## starting httpd help server ...
```

```
##  done
```

Example chi-square for goodness of fit

We are curious if the distribution of fin whales in the bay differs by age group. We have made 5 age groups, juvenile, young adult, adult, mature adult, very old. We have aged 100 fin whales into these 5 age categories. The data is given in "finwhale.age.csv". Read the data in.

```
finwhale.age<-read.csv("finwhale.age.csv")
ages<-table(finwhale.age)
```

We are curious is the animals are distributed evenly across the 5 age groups. This means our null hypothesis is: $H_0 = p_{juvenile} = p_{y.adult} = p_{adult} = p_{m.adult} = p_{old}$

Since we have 5 groups and we are not estimating any parameters our degrees of freedom (df), is equal to $5 - 1 - 4$.

And the alternative hypothesis is that at least one of the probabilities is different. Now, conduct the test using chisq.test in R.

```
chisq.test(ages,p = rep(1/length(ages),length(ages)))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  ages
## X-squared = 0.5, df = 4, p-value = 0.9735
```

Do we reject or accept the null hypothesis, what do we conclude?

Chi-Square test goodness of fit

The test can be used to compare data that you have binned against any distribution. This can be very useful sometimes.

Example chi-square test for independence

We are curious if we find the same sex ratio of leopard sharks in cove or along the beach. We have a dataset that contains the number of leopard sharks found in these two locations and the sex of the leopard shark.

The null hypothesis is that the location of the shark and the sex of the shark are independent from one another. The alternative hypothesis is that the location of the shark and the sex of the shark are not independent from another.

Read in the dataset "leopardshark.csv"

```
leopardshark<-read.csv("leopardshark.csv")
head(leopardshark)
```

```
##   Gender Location
## 1   Male    shore
## 2 Female    shore
## 3   Male    shore
## 4 Female    shore
## 5 Female    shore
## 6   Male    shore
```

```r
my.table<-table(leopardshark)
my.table
```

```
##         Location
## Gender   cove shore
##   Female   43    25
##   Male      7    25
```

Problem Test the hypothesis whether the sex of the leopard sharks is independent of location at .05 significance level.

Solution We apply the chisq.test function to the contingency table. How do you interpret the p-value?

```r
chisq.test(my.table)
```

```
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  my.table
## X-squared = 13.281, df = 1, p-value = 0.0002681
```

The p-value is <0.05, so we reject the null and conclude that there is a statistically significant relationship between gender and location. In otherwords, gender and location are not statistically independent.

Hands On Activity - Chi-Square activity with jellybeans

We will do this in groups. Break into 4 groups.

We have jellybeans and we want to know if the different colors occur with equal probability inside the bag.

- White
- Purple
- Pink
- Red
- Yellow
- Black
- Green
- Orange

**spiced bag has no yellow (just 7 colors)

1. Open the bag, and sort the jellybeans by color. Record the number of jellybeans of each color.
2. Write null and alternative hypotheses.
3. Conduct a Chi-square test to see if all colors are equal.
4. Interpret p-value and draw conclusion.
5. Bonus: Make a barplot of the # of jellybeans of each color in R

- *You may eat the jellybeans once you have collected your data*