The chi-square test

Overview

In general, the chi-square (χ^2) test is a useful statistical test to look at differences with **categorical variables** (e.g., political preference, gender). We can use the χ^2 test in two similar, but subtly difference situations:

1. when estimating how similar an **observed** distribution is to an **expected** distribution (e.g., are the same proportion of people supporting the three political parties?). Here, we use a **goodness-of-fit test** to perform 1-way classification.

2. when estimating whether **two** random variables are **independent/not related** (e.g., is the likelihood of getting aid from large companies vs. small companies the same for all three political parties?). Here we use a **contingency test** to perform 2-way classification.

First we'll compute χ^2 by hand, and then we can do it using R!

1-way classification: Goodness-of-fit test

If you have **one categorical variable** from a single population, and you'd like to determine whether the sample is **consistent** with a **hypothesized distribution**, you can use a χ^2 goodness-of-fit test!

Null hypothesis (H_0): The data follow a specified distribution

Alternate hypothesis (H_a): The data don't follow the specified distribution

χ^2 **Equation**:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E}$$

Here, O_i is the **observed frequency** for bin i and E_i is the **expected frequency** for bin i.

To calculate the **expected frequencies** for bin i:

$$E_i = N * p_i$$

Here, N is the total sample size, and p_i is the hypothesized proportion of observations in bin i. For instance, if you have 3 bins, and you expect the observations to be distributed equally between bins, your proportion p would be $\frac{1}{3}$ for each bin. This test is only really appropriate if you have expected frequencies of at least 1, and hopefully not less than 5 per bin!

Calculating goodness-of-fit by hand

For example, imagine that there are 3 political parties (Democrats, Republicans, and Independent), and you take a poll from a sample of US citizens asking which category they support (i.e., their likely vote). You might wonder whether the people's likely votes are equally distributed between the 3 political parties (your H_0).

First, let's get some data:

```
votes = c(D=587, R=552, I=480)
votes
```

```
##   D   R   I
## 587 552 480
```

Calculate observed proportions

It can be helpful to reframe these observed frequencies as proportions of the total sample. To do that, we just divide each observation by the sample size (N).

```
N = sum(votes); N
```

```
## [1] 1619
```

```
votes_freq = votes/N
votes_freq
```

```
##      D      R      I
## 0.3626 0.3410 0.2965
```

Determine expected frequencies

Now, we want to determine the expected frequencies for each bin. Since we have 3 categories (Democrat, Republican, Independent), and our null hypothesis is that voters are equally distributed across the categories, our **hypothesized proportion** p of voters/bin = $\frac{1}{3}$. Note that we could also have different proportions, e.g., based on data from another year (useful if we wanted to determine if this year's voters are behaving similarly to last year!). We just need to multiply our hypothesized proportion by our sample size, N, to calculate the expected frequencies.

```
p = 1/3; p
```

```
## [1] 0.3333
```

```
E = N * p; E
```

```
## [1] 539.7
```

Since we have 3 categories (Democrat, Republican, Independent), and our null hypothesis is that voters are equally distributed across the categories, our **hypothesized proportion** p of voters/bin = $\frac{1}{3}$. Thus, our **expected frequencies** for each bin should be $N * p$ = 1619 * $\frac{1}{3}$ = 539.6667

	D	R	I
Observed	587	552	480
Expected	539.67	539.67	539.67

Calculate χ^2

Now, using our equation for χ^2:

$$\chi^2 = \frac{(587 - 539.67)^2}{539.67} + \frac{(552 - 539.67)^2}{539.67} + \frac{(480 - 539.67)^2}{539.67} = 11.03$$

```
chisq_rs = (587 - 539.67)^2/539.67 + (552-539.67)^2/539.67 + (480 - 539.67)^2/539.67; ch
isq_rs
```

```
## [1] 11.03
```

Calculate df

$df = k - 1$, where k is the number of categories. So, $df = 3 - 1 = 2$

Calculate significance

```
p_val = pchisq(chisq_rs, df = 2, lower.tail = FALSE)
```

So, we calculated χ^2 (df=2) = 11.03, p = 0.004.

Calculating goodness-of-fit with R

Now, we can do this in one line of code with R!

Calculating with raw observations

```
rs1 = chisq.test(votes); rs1
```

```
##
##  Chi-squared test for given probabilities
##
## data:  votes
## X-squared = 11.03, df = 2, p-value = 0.004025
```

Writing up results

We can reject the null hypothesis that the voters are equally distributed across political parties, χ^2 (df=2) = 11.03, p = 0.004. Thus, the votes are significantly different than we would expect if an equal number of voters ahd voted for each category. From looking at the observed frequencies compared to those expected, it looks like fewer voters were supporting the Independent party (~30%), compared to the Democrats or Republicans (~35%).

Try to give one causal (or otherwise) mechanisms for why this might be the case!

Getting expected values

Here, we can also extract the **expected values** from R! That's the same thing as N/k, where k is the number of categories (in this example, $k = 3$).

```
rs1$expected
```

```
##     D     R     I
## 539.7 539.7 539.7
```

```
N/3
```

```
## [1] 539.7
```

Calculating goodness-of-fit with different proportions

Now, imagine that last election 36% of voters voted for the Democrats, 36% voted for the Republicans, and 28% voted for Independent. We might wonder if this year's votes are significantly different from those expected proportions. To do this with R, we just need to specify one additional parameter, `p`.

```
expected_ps = c(.36, .36, .28)
rs2 = chisq.test(votes, p = expected_ps); rs2
```

```
##
##  Chi-squared test for given probabilities
##
## data:  votes
## X-squared = 3.232, df = 2, p-value = 0.1987
```

It looks like this year's votes are *not* significantly different from the proportions from last election!

2-way classification: Contingency test

If you have **two categorical variables**, and you'd like to determine whether the variables are independent (i.e., that there is no relationship between them), you can use a χ^2 contingency test! This is also sometimes called a **test of independence**.

Null hypothesis (H_0): The 2 categorical variables are independent (there is no relationship between the variables)

Alternate hypothesis (H_a): The 2 categorical variables are dependent (there is a relationship between the variables)

To calculate the **expected value** for a cell:

$$E_{ij} = \frac{R_i C_j}{N}$$

where R = row, C= column, N = total, for ith row and jth column

χ^2 **Equation**:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E}$$

Calculating contingency test by hand

In this example, we have data about juvenile deaths in the County (N=350), specifically about the time of death (`time` = 1, 2, 3; corresponding to 1990-91, 1992-93, 1994-95), and the cause of death (`cause` = maltreatment, or other). We want to know if there is any evidence that, in recent years, we have been seeing an increase in maltreatment cases, many of which result in death? That is, is there reason to think that time and cause of death might be related (dependent)?

We have the cell counts for this table provided, and we can put them into R.

```
death_table = matrix(c(26, 31, 45, 68, 80, 100), ncol=2)

# label the columns and rows
colnames(death_table) = c('maltreat', 'other')
rownames(death_table) = c('1', '2', '3')

# convert to a table
death_table = as.table(death_table)

# add margins to calculate expected values!
addmargins(death_table)
```

```
##     maltreat other Sum
## 1          26    68  94
## 2          31    80 111
## 3          45   100 145
## Sum       102   248 350
```

Calculate expected values

For each cell, we can calculate the expected value by multiplying that cells row and column totals and dividing by our total sample size. For instance, using the margins from above, the expected value for E_{11} (1, maltreat) = $\frac{94*102}{350} = 27.39$. The expected value for E_{12} (2,maltreat) = $\frac{111*102}{350} = 32.35$.

Once we calculate an expected value for each cell, we can then use the same χ^2 function using the observed and expected values to calculate our χ^2 statistic. Try filling in a blank table with the expected values, and then calculate out χ^2!

You should calculate that χ^2 = 0.43

Assess significance:

We can calculate our degrees of freedom using this function:

$$df = (\text{number of rows} - 1) * (\text{number of columns} - 1)$$

Since we have 3 rows and 2 columns, our $df = (3 - 1) * (2 - 1) = 2$

And then use the `pchisq()` used above to determine our p-value/significance!

```
pchisq(0.43, df=2, lower.tail = FALSE)
```

```
## [1] 0.8065
```

We fail to reject the null hypothesis that cause of death is independent of time (i.e., that there is no relationship between the variables), $^2 (df=2) = 0.43, p = 0.81, and thus we don't have any evidence that maltreatment deaths are increasing with time! The causes of deaths for juveniles appears to be pretty stable over time!

Running a contingency test with R

With R, we can calculate all this in one line of code…

```
chisq.test(death_table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  death_table
## X-squared = 0.4308, df = 2, p-value = 0.8062
```

Chi-square tests on data frames

If we don't have the contingency table already, we can create it from a dataframe!

```
df_death = read.csv("http://stanford.edu/class/psych252/data/earlydeaths.csv")

str(df_death)
```

```
## 'data.frame':    350 obs. of  2 variables:
##  $ time : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ cause: Factor w/ 2 levels "maltreat","other": 1 1 1 1 1 1 1 1 1 1 ...
```

```
# Create the table from the 2 categorical variables!
table(df_death)
```

```
##     cause
## time maltreat other
##    1       26    68
##    2       31    80
##    3       45   100
```

```
# Call summary to get chisq
summary(table(df_death))
```

```
## Number of cases in table: 350
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 0.4, df = 2, p-value = 0.8
```

```
# Or use the chisq.test() function
res = chisq.test(table(df_death))
res
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df_death)
## X-squared = 0.4308, df = 2, p-value = 0.8062
```