Lab 9: χ^2 tests

Template for lab report

Write your report, or at least run the code and create the plots, as you go so that if you get errors you can ask your TA to help on the spot. Knit often to more easily determine the source of the error.

```
download.file("http://stat.duke.edu/~cr173/Sta102_Sp14/Lab/lab9.Rmd", destfile = "lab9.Rmd")
```

Goodness of fit test

In November 2, 2011 the Powerball lottery had an estimated jackpot of $245,000,000 causing a rush on tickets here in North Carolina and other Powerball states. The Powerball drawing happens every Wednesday and Saturday at 10:59pm, and since 2006 when North Carolina started participating in the Powerball there have been 1,460 drawings as of November 2, 2011.

In the basic Powerball game, players select 5 numbers from a set of 59 white balls, and 1 number from 39 red Powerballs. “In each drawing, winning numbers are selected using two ball machines; one contains white balls numbered 1 through 59; the other contains the red Powerballs numbered 1 through 39. Five balls are drawn from the first machine, and one from the second machine; these are the winning numbers. Games matching at least three white balls and/or the red Powerball win.”[1] The draws are made without replacement.

We will first focus on the 59 white balls and the frequencies with which they have been drawn.

> **Exercise 1** If each ball is equally likely to be drawn, how many times would you expect each ball to have been drawn so far?

The table below shows the drawing frequencies of the 59 white balls in the 1,460 games played so far.[2]

ball	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
freq	125	136	127	132	143	129	137	138	138	138	129	147	141	139	139

ball	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
freq	152	135	130	145	159	124	150	131	131	114	155	139	130	116	144

ball	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
freq	131	159	118	132	140	135	142	137	140	151	158	154	141	127	146

ball	46	47	48	49	50	51	52	53	54	55	56	57	58	59	
freq	122	132	146	146	94	83	89	78	57	55	21	21	26	26	

Let’s load these data into RStudio. Load the powerball.R file from the course website using the source function.

```
source("http://stat.duke.edu/~cr173/Sta102_Sp14/Lab/powerball.R")
```

Your workspace should have three new variables: `freq_white_full`, `freq_white`, and `freq_red`. The first variable contains all 1460 draws while the other two contain 295 recent white and red draws respectively.

[1] From Wikipedia: *http://en.wikipedia.org/wiki/Powerball*.

[2] From *http://www.us-lotteries.com/north_carolina/powerball/powerball-numbers-analysis.asp*.

Exercise 2 Create a plot to visualize the distribution of draw frequencies of freq_white_full. Add a horizontal red line to the bar plot indicating the expected frequency for each ball (use the abline function). (Note that this data reflects counts and as such it is not appropriate to use a histogram)

Exercise 3 Do you see anything unexpected in the plot? Can you guess what the reason for this might be?

Due to this discrepancy we'll focus on the last 295 draws prior to November 2nd, 2011 which are contained in freq_white and freq_red.

Exercise 4 Make a plot of the frequencies of both white and red draws. For each plot be sure to draw a horizontal red line indicating the expected frequency for each ball type. Note that this value will be different than the value computed earlier since we're looking at a fewer number of drawings. Do you see anything unexpected in the bar plot?

We'll now perform a chi-square test on both red and white frequencies to see if the observed frequencies follow the distribution of the expected frequencies.

Exercise 5 Write the appropriate hypotheses for testing if the draws are fair.

An example with dice

We will now take a brief detour and work through an example of how to perform the steps necessary to perform a goodness of fit hypothesis test. We will do this by generating simulated die rolls and then calculating the quantities needed for the chi-square test statistic and p-value.

```
set.seed(1234)
# generate die rolls
(die = sample(1:6, 125, replace = TRUE))

##   [1] 1 4 4 4 6 4 1 2 4 4 5 4 2 6 2 6 2 2 2 2 2 2 1 1 2 5 4 6 5 1 3 2 2 4 2 5 2 2 6 5 4 4
##  [43] 2 4 2 4 5 3 2 5 1 2 5 4 1 4 3 5 2 6 6 1 2 1 2 5 2 4 1 4 1 6 1 5 1 4 3 1 2 5 6 3 1 4
##  [85] 2 6 3 2 1 6 1 6 1 1 1 4 2 1 2 5 1 4 2 2 1 2 1 1 3 1 5 1 6 1 2 6 6 2 1 5 5 6 6 6 3

# generate frequency table of die rolls
(freq_die = table(die))

## die
##  1  2  3  4  5  6
## 29 33  8 21 16 18

# entries in the tables are counts, summing them gives the total die rolls
(n = sum(freq_die))

## [1] 125

# expected number of each roll
(E = n * 1/6)

## [1] 20.83

# observed number of each roll
(O = c(freq_die))

##  1  2  3  4  5  6
## 29 33  8 21 16 18
```

```
E - O

##        1        2        3        4        5        6
##  -8.1667 -12.1667  12.8333  -0.1667   4.8333   2.8333

(E - O)^2

##         1         2         3         4         5         6
##  66.69444 148.02778 164.69444   0.02778  23.36111   8.02778

(E - O)^2/E

##        1        2        3        4        5        6
## 3.201333 7.105333 7.905333 0.001333 1.121333 0.385333

(chisq = sum((E - O)^2/E))

## [1] 19.72

# calculate df
df = 6 - 1

# calculate upper tail probability (p-value = P(>chisq))
pchisq(chisq, df = df, lower.tail = FALSE)

## [1] 0.00141
```

Exercise 6 After running the above code would you conclude the die is fair? Has an error been made here (based on the hypothesis test), if so what kind of error?

Additional Exercises

Using the approach demonstrated on the dice you will now conduct hypothesis tests to evaluate if the draws of white and red balls in the Powerball lottery are random. Be sure to indicate whether necessary conditions are met and discuss the implications of your conclusions. Note for the following exercises (7 and 8) you must explicitly calculate the Chi-square test statistic and p-value.

Exercise 7 Conduct a hypothesis test to evaluate if the white ball draws are fair (uniformly distributed).

Exercise 8 Conduct a hypothesis test to evaluate if the red ball draws are fair (uniformly distributed)..

Exercise 9 Based on your results to the proceeding questions, how would you choose to play the Powerball lottery?

R also has a built in function for performing a chi square test, conveniently it is named `chisq.test`. If the function is passed a one dimensional frequency table it will automatically perform a goodness of fit test. Again using our simulated die rolling data we see that the results of the `chisq.test` function matches are results above.

```
chisq.test(freq_die)
```

Exercise 10 Use the `chisq.test` function to check your results to Exercises 7 and 8.