Permutation chi-square test

Solutions

1. Our hypotheses are:

$$H_0 : \text{Sports preference and gender are independent.}$$
$$H_1 : \text{Sports preference and gender are dependent.}$$

2. It would not be appropriate to perform a traditional chi-square test in this case, because we have expected counts of less than 5 in <u>each</u> of the cells of the table.

3. Assuming we still have 2/5 prefer basketball, 2/5 prefer football, and 1/5 prefer other, and that we have 3/5 boys and 2/5 girls, we can figure out how many students we would have to survey for the traditional chi-square test to be valid. The limiting table cell will be the girls/other cell, since the other is the least represented preference, and the girls are the least represented gender.

   The expected count in the girls/other cell will be 2/5*(# others), and we want this to be at least 5. So we solve:
$$2/5 \cdot (\text{\# others}) \geq 5$$
   Which gives us:
$$\text{\# others} \geq 12.5$$
   Since the # other is just 1/5 of the total number of students, we need to survey 5*12.5 = 62.5 students. So we need to survey at least 63 students.

4. Our observed chi-square statistic is 2.9167. (Note that the p-value given by `chisq.test()` may not be valid, because the chi-square approximation may not be valid.)

```
> Row1 = c(1,2,0); Row2 = c(1,0,1)
> Table = rbind(Row1, Row2)
> chisq.test(Table)

        Pearson's Chi-squared test

data:  Table
X-squared = 2.9167, df = 2, p-value = 0.2326

Warning message:
In chisq.test(Table) : Chi-squared approximation may be incorrect
```

5. There will be `choose(5,2) = 10` rows in our table.

boys			girls	
B_1	B_2	F_1	F_2	O_1
B_1	B_2	F_2	O_1	F_1
B_1	B_2	O_1	F_1	F_2
F_1	F_2	B_1	B_2	O_1
F_1	F_2	B_2	O_1	B_1
F_1	F_2	O_1	B_1	B_2
F_1	B_1	O_1	F_2	B_2
F_2	B_2	O_1	F_1	B_1
F_1	B_2	O_1	B_1	F_2
F_2	B_1	O_1	B_2	F_1

6. Our table with contingency table assignments is below:

boys			girls		
B	F	O	B	F	O
2	1	0	0	1	1
2	1	0	0	1	1
2	0	1	0	2	0
1	2	0	1	0	1
1	2	0	1	0	1
0	2	1	2	0	0
1	1	1	1	1	0
1	1	1	1	1	0
1	1	1	1	1	0
1	1	1	1	1	0

7. Our permutation distribution is in the table below:

boys			girls			test statistic (X^2)
B	F	O	B	F	O	
2	1	0	0	1	1	2.9167
2	1	0	0	1	1	2.9167
2	0	1	0	2	0	5
1	2	0	1	0	1	2.9167
1	2	0	1	0	1	2.9167
0	2	1	2	0	0	5
1	1	1	1	1	0	0.8333
1	1	1	1	1	0	0.8333
1	1	1	1	1	0	0.8333
1	1	1	1	1	0	0.8333

```
> Row1 = c(2,1,0); Row2 = c(0,1,1); Table = rbind(Row1,Row2)
> chisq.test(Table)

        Pearson's Chi-squared test

data:  Table
X-squared = 2.9167, df = 2, p-value = 0.2326

Warning message:
In chisq.test(Table) : Chi-squared approximation may be incorrect
```

8. The p-valeu for this test is $6/10 = .6$. (This is because 6 of the 10 permutations resulted in a test statistic as or more extreme than the one we observed!) We have no evidence that gender and sports preference are dependent.

9. Our test statistic is $X^2 = 8.744$ and our p-value is 0.015. We have evidence that gender and sports preference are related.

```
### Make observed contingency table and calculate stat
Row1 = c(22,40,12); Row2 = c(30,17,11)
Table = rbind(Row1,Row2)
teststat.obs = chisq.test(Table)$statistic
teststat.obs

### create the prefernce data and the gender data
preference = c( rep("B",52), rep("F",57), rep("O",23))
gender = c( rep("boy",74), rep("girl",58) )
table(preference); table(gender)

y = preference; x = gender
teststat = rep(NA, 1000)

for(i in 1:1000) {

### randomly "shuffle" the y data between the x groups
ySHUFFLE = sample(y)

### compute chi-square stat for the shuffled data
TableSHUFFLE = table(x,ySHUFFLE)
teststat[i] = chisq.test(TableSHUFFLE)$statistic
}

### calculate the approximate p-value
sum(teststat >= teststat.obs)/1000

> teststat.obs
X-squared
 8.744026
> sum(teststat >= teststat.obs)/1000
[1] 0.015
```

10. Yes, we could have done a traditional chi-square test here, since our expected counts are all greater than 5. A traditional chi-square test gives us the same test statistic, and we get a p-value of 0.013, which is pretty close to our permutation p-value. We still have evidence that gender and sports preference are related.

```
> Row1 = c(22,40,12); Row2 = c(30,17,11)
> Table = rbind(Row1,Row2)
> chisq.test(Table)

        Pearson's Chi-squared test

data:  Table
X-squared = 8.744, df = 2, p-value = 0.01263
```