Permutation chi-square test

Exact permutation chi-square test

Suppose we asked 5 middle school students (3 boys and 2 girls) which sport they prefer – football, basketball, or something else. They results are summarized in the table below:

	basketball	football	other
boys	1	2	0
girls	1	0	1

We want to perform a test to determine whether gender is related to preferred sport.

1. List the appropriate null and alternative hypotheses for this test.

2. Would it be appropriate to perform a traditional chi-square test in this situation? Why or why not?

3. Assuming that the proportion of middle school students who prefer each sport stays the same (for instance, that 2/5 still prefer basketball, etc), how many students would we have to survey for the traditional chi-square test to be valid?

Instead of the traditional chi-square test, you will be performing a permutation chi-square test BY HAND to do this analysis. We start by calculating the observed test statistic (X^2_{obs}) for our data:

4. Calculate the chi-square statistic for the observed data. (You can do this by entering the data into `R` as a table, and then using `chisq.test()`)

To do the permutation test, we want to list out every possible way that these 6 sports prefences (B_1, B_2 for basketball, F_1, F_2 for football, and O_1 for other) can be divided between the two genders. We do this because *if the null hypothesis is true*, then the preferred sport is *independent* of gender, so the preferred sports can be assigned to *any* of the genders. We just need to be sure that we keep the number in each preferred sport and the number in each gender the same. We can do this by creating a table similar to the one below:

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

5. How many rows will be in our table (how many assignments are possible)? Create a table and fill in all of the possible assignments.

Next we need to create a contingency table for each of these preferred sports assigments. We will do this by creating a second table with contingency table assignments. The first possible assignment has 2 boys who prefer basketball, 1 boy who prefers football, 1 girl who prefers football, and 1 girls who prefers something else. The second row of assignments will have this same contingency table.

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

6. Create a table and fill in all of these possible contingency table assignments.

Now, we need to calculate the test statistic (X^2) for each of these contingency table assignments. We can do this using `R`. For the first assignment, you could use the following code:

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

We can then add a column to our table which lists the test statistic for that possible assignment.

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
⋮	⋮	⋮	⋮	⋮	⋮	⋮

7. Add a column to your table for the test statistic, and fill in the value of the test statistic for each possible assignment.

Once you have done this, you have defined the permutation distribution of the test statistic, X^2, for this data set. Now you can use this permutation distribution to calculate a p-value.

8. Calculate the p-value for this permutation test. Also state a conclusion, in context, about the result of this test.

Random sampling the permutations in `R`

In general, we will let `R` create the permutation distribution for us, although you should be able to do so by hand in a small data set. Suppose now we have the full data on the survey of middle school students:

	basketball	football	other
boys	22	40	12
girls	30	17	11

The code below will run a permutation chi-square test for this data set. It follows the same steps we did by hand. One important thing to notice that is different from pervious permutation tests we have done: when we create the preference and gender data, you can think of this as creating B_1, B_2, . . . B_{52}, F_1, F_2, . . . , F_{57}, O_1, O_2, . . . O_{23} so we can permute them between the genders.

```
### 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
```

9. Run the permutation test. Report your test statistic and your p-value. Also state a conclusion, in context, about the result of the test.

10. Could you have used a traditional chi-square test here? Why or why not? Perform a traditional chi-square analysis and compare to your permutation result.

Lab Summary

Complete your Lab Summary in a separate document. Please write clearly and in complete sentences. Do not include any R code in your summary.

1. Summarize the results from your analysis of the larger sports preference data set. Include your hypotheses (in words), your test-statistic, and your p-values from both the traditional and permutation tests. Also include histogram of your permutation distribution with the p-value shaded. Then state a conclusion in the context of the original problem. Write up your summary in a neat little paragraph, and include the graph at the end. Do not include any R code in your summary.

2. Suppose the 'Other' category could be broken down into the following sports: swimming, soccer, and track. Create a data table with football, basketball, and these three additional sports that preserves the original data but where you would not feel comfortable performing a traditional chi-square test.