Lab 7 - Chi-squared Test for Independence

North Carolina Births

In 2004, the state of North Carolina released to the public a large data set containing information on births recorded in this state. This data set has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children. We will work with a sample of observations from this data set. These cases were chosen at random.

We have observations on 13 different variables, some categorical and some numerical. The meaning of each variable is as follows.

`fage`	father's age in years.
`mage`	mother's age in years.
`mature`	maturity status of mother.
`weeks`	length of pregnancy in weeks.
`premie`	whether the birth was classified as premature (premie) or full-term.
`visits`	number of hospital visits during pregnancy.
`gained`	weight gained by mother during pregnancy in pounds.
`weight`	weight of the baby at birth in pounds.
`lowbirthweight`	whether baby was classified as low birthweight ('low') or not ('not low').
`gender`	gender of the baby, 'female' or 'male'.
`habit`	status of the mother as a 'nonsmoker' or a 'smoker'.
`marital`	whether mother is 'married' or 'not married' at birth.
`whitemom`	whether mom is 'white' or 'not white'.

Let's load the `nc` data set into our workspace:

```
nc <- read.csv("nc.csv", header=T)
attach(nc)
```

We will first tackle the relationship between a mother's smoking habit and the health of her baby. By now you have had practice using R commands to summarize and visualize data.

Exercise 1: What proportion of total births were to mothers that were smokers? What proportion of total births were of babies that were classified as low birthweight?

Exercise 2: Make a contingency table of `lowbirthweight` vs. `habit` and then make a mosaic plot. What does the plot tell us about the relationship between the two variables?

Exploratory analysis is a useful first step when dealing with data because it helps us notice trends and develop research questions. The mosaic plot we just made shows us that in our sample, the babies born to smokers were more often classified as low birthweight than those born to non-smokers.

We can quantify this observed difference as the difference between the proportions of low birthweight babies born to mothers who smoke and do not smoke.

Exercise 3: Calculate the proportions of low birthweight babies born to non-smoker and smoker mothers, and find the difference between the two proportions. Take note of this value as it will be useful in the next stage of the analysis.

At this stage, though, that's just a descriptive statistic. In order to determine if this difference is significant and not due to chance, we need to do inference.

Let's conduct a hypothesis test to answer the following question: is smoking and having a low birthweight related (dependent)?

Exercise 4: State the hypotheses for this research question in words and using probabilities.

To test these hypotheses, we can use a χ^2 test or a randomization based simulation method.

Independence test

The function `chisq.test()` implements the classical χ^2 test for independence. Additionally, the `chisq.test()` function also allows for a simulation based approach to be used when the assumptions for the classical test are not met.

Exercise 5: Does the birth data meet the assumptions necessary for the χ^2 test to be valid?

Run the following command to execute the χ^2 test

```
chisq.test(table(habit,lowbirthweight),correct=FALSE)
```

If the cell counts are not sufficiently large, use `chisq.test(table(habit,lowbirthweight),correct=FALSE, simulate.p.value = TRUE)` to test for an association. By default R will use 2000 simulations, if you would like to use more you can specify the number with the `B` function argument. Hint: If the counts are not large enough to use a χ^2 distribution, R will warn you with the message `Chi-squared approximation may be incorrect` at the bottom of the output. If you get this message, you should use the p-value based on simulation instead.

Exercise 6: What is the p-value? Do these data provide convincing evidence that your two variables are associated?

On Your Own

We've considered the association between smoking and low birthweight. Now it's your turn to analyze pairs of variables that are of interest to you.

1. Select two categorical variables from this data set that is of interest to you. You will be performing a hypothesis test that compares the dependence of these two variables. What is the research question you will be answering? What are the null and alternative hypotheses?
2. Conduct an appropriate hypothesis test to evaluate your hypotheses. If the classical test is appropriate you should run use both the classical test and the simulation test, otherwise just use the simulation test. Include the code you used to run the test(s) and any output that results. Based on these results what can you conclude about your hypotheses? If you ran both tests do the results agree with one another?
3. Repeat parts 1. and 2. for two additional pairs of categorical variables.