­Chi-squared – Lab

**Introduction:**

If you run into trouble with any of these commands, R has a large user community that frequently posts and responds to technical questions. Searching online can usually help find answers by searching for the command you are unsure about.

**Part 1**

**An association between age at first pregnancy and cervical cancer?**

In medical research, retrospective case-control studies are used to identify correlations between a possible predictive variable and an outcome. These studies work by selecting subjects who have already been diagnosed with a specific disease and then selecting a separate set of similar subjects who haven’t been diagnosed with the disease. The table below shows the results of a case-control study of risk factors for cervical cancer (Journal of the National Cancer Institute, 1979; 63:23).

|  |  |  |
| --- | --- | --- |
| Age at first pregnancy | Cervical cancer | No cervical cancer |
| ≤25 | 42 | 203 |
| >25 | 7 | 114 |

The table above is called a 2x2 contingency table, and it compares 2 categorical variables against two other categorical variables (known as factors in R). It has a total of four cells.

a) First, open RStudio and do a chi-squared test of independence.

To do this, you can copy and paste the command below into the console

chisq.test(rbind(c(42,203),c(7,114)),correct)

The first result you get does a chi-squared test of independence using the upper row and then the lower row in order. Now, switch the ordering of your chi-squared test so that the chisq.test formula uses the left column as the first input parameter, and then the right column as the second input parameter. Report both p-values and comment.

test 1 p-value: 0.002683975

test 2 p-value: 0.002683975

P-values are the same because the marginals are the same in regards to each pair of numbers however the placement of those numbers changes. The calculations remain the same.

In order to understand what is going on you need to understand what a null hypothesis is, which we will explore in the next sections.

b) In R, calculate the total number of women, as well as the fraction of women who had cervical cancer (**1.**) and those who didn’t (**2.**). Also calculate the fraction of women who had their first pregnancy before or after 25 years (**3. and 4.**). Add this information to the appropriate cells below. These are called marginal probabilities.

|  |  |  |  |
| --- | --- | --- | --- |
| Age at first pregnancy | Cervical cancer | No cervical cancer |  |
| ≤25 | 42 | 203 | **3. 245** |
| >25 | 7 | 114 | **4. 121** |
|  | **1. 49** | **2. 317** | **Total = 366** |

c) Now, we will assume a null hypothesis, which is an idea about the data that we are going to test and (hopefully reject) with statistics.

The null hypothesis is that there is no association between cervical cancer and the age at first pregnancy. In other words, the probability of being diagnosed with cervical cancer is not dependent on the age of first pregnancy. **Recall that if A and B are independent, then P(AB)=P(A)P(B).**

Armed with this knowledge, calculate the probability of any particular woman being in each of the four categories below, under the null hypothesis. Check that the probabilities sum to 1.

|  |  |  |
| --- | --- | --- |
| Age at first pregnancy | Cervical cancer | No cervical cancer |
| ≤25 | .0896 | .5797 |
| >25 | .0442 | .2863 |

Now, show the expected value of women in each of the four categories below. Note: these are the **expected values**, E.

|  |  |  |
| --- | --- | --- |
| Age at first pregnancy | Cervical cancer | No cervical cancer |
| ≤25 | 32.8 | 16.2 |
| >25 | 212.2 | 104.8 |

Use R, to calculate the value of the test statistic with the following formula:

X2 = 8.057873

where ‘actual’ refers to the real, observed data and ‘expected’ refers to the expected values you obtained above. What is the p-value calculated from R?

p value = .00453

d) We can simulate what would happen if we observed this process many times, under the null hypothesis of no association, and with fixed marginal probabilities (total number of women with and without cancer, total number of women who had a pregnancy before or after age 25)

Say that in our first simulation, we obtained 49 women who had cervical cancer and had a baby before 25. Fill in the other three boxes.

|  |  |  |
| --- | --- | --- |
| Age at first pregnancy | Cervical cancer | No cervical cancer |
| ≤25 | 49 | 196 |
| >25 | 0 | 121 |

The point here is that once you know one of the boxes above in a 2x2 contingency table, you can fill in the remaining entries for the whole table, given the marginal probabilities.

If instead you had a 3x4 contingency table, after knowing the 7 marginal probabilities, how many boxes would you have to fill in order to determine the whole table?

2 boxes

This is an example of a 3x4 contingency table:

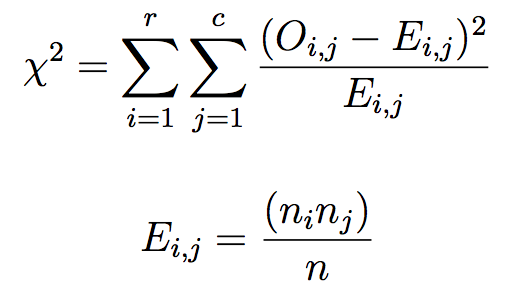
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Diet | Colon cancer | Stomach cancer | Lung cancer | No cancer |
| Vegetarian | \* |  |  |  |
| Bacon and hotdogs | \*\* |  |  |  |
| Chicken and Fish |  |  |  |  |

Degrees of Freedom:

df = (r-1)(c-1) = (2\*3) = 6

This number is called the degrees of freedom (df). Our 2X2 table has 1 degree of freedom and the 3x4 table has df equal to your answer above.

e) Next, we will compute the chi-squared statistic. A statistic is a function of the data that reflects something about the data set. In our case, the statistic will measure deviations from the expected values calculated in (c).



The chi-squaredd statistic is the sum of the squared deviations from expectation in each cell. *Oi,j* are the four Observed values, *Ei,j* are the expected values and *r* is the number of rows, and *c* is the number of columns (in the contingency table).

Compute the chi-squared statistic for the actual data (shown at the beginning of the lab) in R.

26.2

Now compute the statistic for the simulation below:

|  |  |  |
| --- | --- | --- |
| Age at first pregnancy | Cervical cancer | No cervical cancer |
| ≤25 | 33 | 212 |
| >25 | 16 | 105 |

0.004235649

As you can see, the bigger the difference between that data and the expectation, the larger the chi-squared statistic gets.

f) Now that you have a null hypothesis and a statistic, you need to know how this statistic behaves under the null hypothesis.

To do this, imagine if we went to a village in which the age at first pregnancy is independent of cervical cancer. In that village, if we took many random samples of 366 women in which 49 had cancer and the rest had no cancer, and 245 had a baby before age 25, we could calculate the chi-squared statistic for each of those random samples. If we did this we would see this statistic (which is a random variable) which would have a distribution.

To discover this distribution you could do two things: we could run a simulation, or we could statistically analyze the distribution. In this exercise we can tell you that the statistic follows a chi-squared distribution with 1 degree of freedom (assuming that all the cells in your table are at least 5).

We will now use R to visualize this chi-squared distribution.

First obtain a random sample of size 20,000 from this distribution by pasting the following command in R:

chi2 <- rchisq(20000,df=1)

Now you can see a histogram of the data with:

hist(chi2,30)

To add a line where the data’s chi-squared statistic is:

abline(v=9.01)

Now you can try to see what happens when you vary the df parameter, or the bin number in the hist() command.

We could also make the histogram in Excel, but we don’t need to anymore. Instead we will use a function in R to do this for us automatically.

g) Estimate the probability that a chi-squared df=1 random variable takes a value greater than 9.01.

To do this, you might use a command similar to the one below which gives you the number of values in chi2 that are greater than 5:

length(chi2[chi2 > 5])

chi[chi>5] gives you a list of values which are greater than 5, length() tells you the size of that list.

Explain the relationship between the number you just obtained and the p-value in (a)?

471 values > 5 : P(values > 5) = 471/20000 = .02545

p value shows the shows a significance level greater than alpha in our case > .05

In brief, a p-value is the probability that you would get a statistic that is more extreme than the one obtained in your data, if the null model were true.

h) As you got a p-value < 0.01, you can reject the null hypothesis at the 0.01 level.

Can you say that you found conclusively that pregnancy before age 25 causes cervical cancer?

Our initial p-value was 0.002 < 0.01 which means we reject H0

We can say that pregnancy age and cancer are non-independent of each other. We can't say for certain one is a direct causation of another.

i) In practice, if some of the cells in your contingency table are small, the chi-squaredd test used above is not the best choice, as the chi-squared statistic computed under the null model will not behave exactly as predicted. A better choice in these circumstances is the Fisher exact test. It is exact because it gives us precise a p-value without making many assumptions about the distribution of statistics.

Run this test, which calculates the Fisher exact test on the same dataset in R and comment on the results:

fisher.test(rbind(c(42,203),c(7,114)))

p-value = 0.002937

results are very close to that of Pearson's p-value

Wrapping up: so far, you have run a particular set of statistical tests, but the philosophy behind all other tests you will run is very similar to what you just did:

1. Come up with a null hypothesis
2. Select the test statistic
3. Select the level of significance
4. Compute a statistic from the data
5. Reject or fail to reject null hypothesis

**Part 2**

Assume we manufacture metal parts using a computer controlled milling machine. We want to compare the work of two different mill operators who are creating many of the same part. We will promote the mill operator with the lower variance in his parts, if the difference is statistically significant. We choose a critical dimension and measure it on 50 parts for mill operator 1 and mill operator 2. Mill operator 1 has a standard deviation of 3.63 units on that measurement, and Mill operator 2 has a standard deviation of of 3.42 units on that measurement. We want to test a hypothesis to prove with *p* < 0.05 that Mill operator 2 has a lower variance than Mill operator 1. Can we promote Mill operator 2? Why or why not?

You can use pf() in R, try using the command:

?pf

to look at the documentation about how to use pf­­­­­

Can't reject the null hypothesis, thus you can not say that Mill 2 has a variance that is significantly less than Mill 1

**Part 3**

The comic XKCD published the following cartoon below. This is known as the multiple testing problem. Why is the link between green jelly beans and acne problematic?

