15.8.1 Category Complexities

Now that Jeremy has learned about statistical concepts such as t-test and linear regression, he feels comfortable analyzing numerical datasets. However, Jeremy recognizes that not all data will be numerical, and eventually he will need to analyze data that is completely categorical. Once again, Jeremy must go back and learn another statistical test that will enable him to compare and contrast the frequency of categorical data.

As we learned previously, categorical data is generally any data that is not measured, or qualitative data. Even though categorical data may not require an instrument to measure, it can be just as informative as numerical data.

One common form of categorical data is **frequency data**, where we record how often something was observed within a single variable. For example, in the mpg dataset, if we were to count up the number of vehicles for each vehicle class, the output would be a form of frequency data.

In data science, we'll often compare frequency data across another dichotomous factor such as gender, A/B groups, member/non-member, and so on. In these cases, we may ask ourselves, "Is there a difference in frequency between our first and second groups?" To test this question, we can perform a chi-squared test.

The **chi-squared test** is used to compare the distribution of frequencies across two groups and tests the following hypotheses:

H_0 : There **is no difference** in frequency distribution between both groups.

H_a : There **is a difference** in frequency distribution between both groups

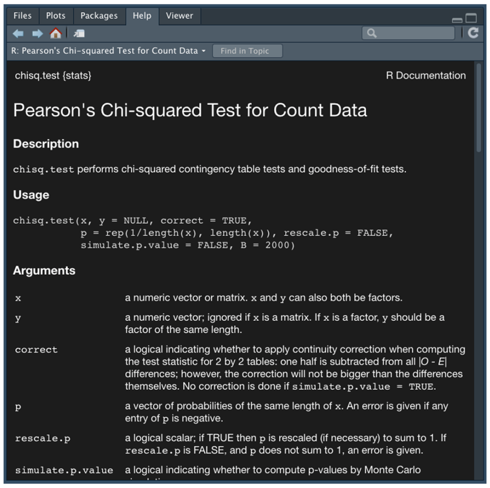
Before we can perform our chi-squared analysis, we must ensure that our dataset meets the assumptions of the statistical test:

1. Each subject within a group contributes to only one frequency. In other words, the sum of all frequencies equals the total number of subjects in a dataset.
2. Each unique value has an equal probability of being observed.
3. There is a minimum of five observed instances for every unique value for a 2x2 chi-squared table.
4. For a larger chi-squared table, there is at least one observation for every unique value and at least 80% of all unique values have five or more observations.

Once we have confirmed our categorical dataset meets all of the assumptions of the chi-square analysis, we can perform our chi-squared test.

In R, we'll compute our chi-squared test using the `chisq.test()` function. Type the following code into the R console to look at the `chisq.test()` documentation in the Help pane.

```
>?chisq.test()
```



chisq.test {stats} R Documentation

Pearson's Chi-squared Test for Count Data

Description

`chisq.test` performs chi-squared contingency table tests and goodness-of-fit tests.

Usage

```
chisq.test(x, y = NULL, correct = TRUE,
           p = rep(1/length(x), length(x)), rescale.p = FALSE,
           simulate.p.value = FALSE, B = 2000)
```

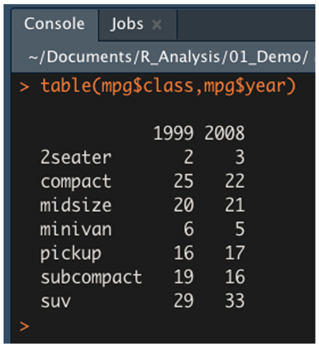
Arguments

`x`	a numeric vector or matrix. `x` and `y` can also both be factors.
`y`	a numeric vector; ignored if `x` is a matrix. If `x` is a factor, `y` should be a factor of the same length.
`correct`	a logical indicating whether to apply continuity correction when computing the test statistic for 2 by 2 tables: one half is subtracted from all $\|O - E\|$ differences; however, the correction will not be bigger than the differences themselves. No correction is done if `simulate.p.value = TRUE`.
`p`	a vector of probabilities of the same length of `x`. An error is given if any entry of `p` is negative.
`rescale.p`	a logical scalar; if TRUE then `p` is rescaled (if necessary) to sum to 1. If `rescale.p` is FALSE, and `p` does not sum to 1, an error is given.
`simulate.p.value`	a logical indicating whether to compute p-values by Monte Carlo

Depending on the structure of your dataset, you can implement the `chisq.test()` function in multiple ways using the optional arguments. The most straightforward implementation of `chisq.test()` function is passing the function to a contingency table. A **contingency table** is another name for a frequency table produced using R's `table()` function. R's `table()` function does all the heavy lifting for us by calculating frequencies across factors.

For example, if we want to test whether there is a statistical difference in the distributions of vehicle class across 1999 and 2008 from our mpg dataset, we would first need to build our contingency table as follows:

```
> table(mpg$class,mpg$year) #generate contingency table
```

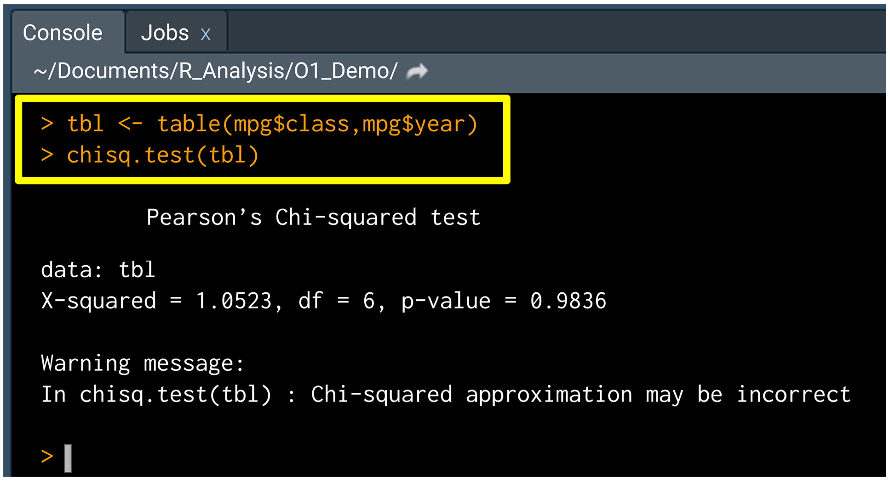


```
> table(mpg$class,mpg$year)

             1999 2008
  2seater       2    3
  compact      25   22
  midsize      20   21
  minivan       6    5
  pickup       16   17
  subcompact   19   16
  suv          29   33
>
```

Then, pass the contingency table to the `chisq.test()`function:

```
> tbl <- table(mpg$class,mpg$year) #generate contingency table
> chisq.test(tbl) #compare categorical distributions
```



```
> tbl <- table(mpg$class,mpg$year)
> chisq.test(tbl)

        Pearson’s Chi-squared test

data:  tbl
X-squared = 1.0523, df = 6, p-value = 0.9836

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

>
```

IMPORTANT

> The chi-squared warning message is due to the small sample size. Because the p-value is so large, we are not too concerned that our interpretation may be incorrect.

Despite having no quantitative input, the chi-squared test enables data scientists to quantify the distribution of categorical variables. Although this test can be applied to more groups and larger datasets, it does have a limit. Increasing the number of groups also increases the likelihood that insignificant changes will incorrectly be considered significant. Therefore, it's important to keep the number of unique values and groups relatively low. A good rule of thumb is to keep the number of unique values and groups lower than 20, which means the degrees of freedom (`df` in the output) is less than or equal to 19.

Take some additional time to practice implementing contingency tables and chi-squared tests using categorical data from our previous datasets. Feel free to tweak the frequency values in the contingency tables tp see what happens to the chi-squared and p-value metrics.

© 2020 - 2022 Trilogy Education Services, a 2U, Inc. brand. All Rights Reserved.