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:

H0 : There **is no difference** in frequency distribution between both groups.

Ha : 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.

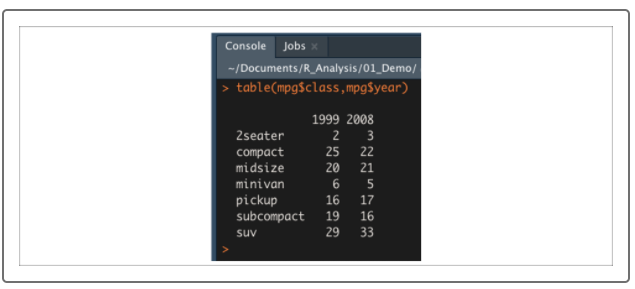
>?chisq.test()



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



> tbl <- table(mpg$class,mpg$year) #generate contingency table

> chisq.test(tbl) #compare categorical distributions

