Testing for an association between two categorical variables

Overview:

We often want to test whether there is any statistical association between two categorical variables. For example, we may collect data on a group of 170 patients who were treated with a specific regimen. We may hypothesize that older individuals (65+ years old) are more likely to experience a mild adverse event (low blood sugar) due to treatment compared to younger individuals (less than 65 years old).

The data:

	Low blood sugar	
Age	**Yes**	**No**
< 65	19	51
65+	65	35

We assume that these data are **independent** as each observation in the table is assumed to be unrelated to any other observation.

An example of an association analysis with **dependent** variables is if we examine adverse events in a set of randomly sampled individuals at two different time points (pre and post treatment). For example:

		Low blood sugar post treatment	
Low blood sugar pre-treatment		**Yes**	**No**
	Yes	60	24
	No	16	70

Do you see the difference in the table structure? Now that we have individuals measured at two different time points, there is a natural pairing between observations from the same individual. Someone who typically has low blood sugar pre-treatment is more likely to have low blood sugar post treatment. And someone who typically has normal blood sugar is more likely to have normal blood sugar post-treatment.

The focus is therefore now on the *off-diagonals* of the above table. Individuals who have no change in low blood sugar status pre to post treatment are not of interest and have no bearing on analysis results. If treatment is not associated with low blood sugar, then we would expect similar proportions to have low blood sugar pre, but not post, treatment as we would to have low blood sugar post, but not pre, treatment.

As usual, statistical tests (such as McNemar's test) to specifically account for this type of dependency are required with clustered data.

What statistical test do we use?

The form of your data dictates the type of statistical test you need. Because we have two categorical variables, to test for association we mainly use **Chi-square tests**. If sample sizes are small (counts < 5 in one or more table cell), **Fisher's Exact test** is a more valid approach and should be used instead. If we have ordinal data, then **Cochran Armitage test** for trend is also useful. Finally, if we have dependent data, **McNemar's test** is utilized.

Where can I read more about these tests?

There are many **excellent online resources** that will allow you to more fully understand how each statistical test is computed and the assumptions necessary to obtain valid results.

How do I perform these tests?

There are plenty of statistical packages and online tools that can be used to test for group differences. Here are some **basic instructions** for performing Chi-square tests, Fisher's Exact tests, and Cochran Armitage trend tests using the R software package.