Testing associations between two categorical variables: chi-square tests

Example: We have a dataset called data1, which consists of 146 observations (patients) and 5 variables (id, treat, age, sex and wt). Participants have been randomized to receive an active treatment or placebo. We will only consider three baseline characteristics: age, sex and weight.

```
head(data1)          #prints the first few rows of the data set
```

id	treat	age	sex	wt
1	Treated	18	Female	62.6
2	Treated	50	Male	57.4
3	Treated	37	Male	104.6
4	Treated	25	Female	55.5
5	Placebo	60	Female	58.4
6	Treated	44	Female	41.9

Suppose we would like to examine whether there is an association between treatment group and biological sex. As patients are randomized to treatment group, we do not expect there to be an association.

We first can construct the 2x2 table for age and treatment group:

```
table(data1$treat, data1$sex)
```

Which yields the following 2x2 table:

	Female	Male
Placebo	34	18
Treated	53	41

To perform a Chi-square test for association:

```
chisq.test(data1$treat,data1$sex)
```

The R output is below:

```
Pearson's Chi-squared test with Yates' continuity correction
data:  data1$treat and data1$sex
X-squared = 0.78376, df = 1, p-value = 0.376
```

The p-value is 0.376, which means that there is no statistically significant association between treatment group assigned and biological sex.

***

If we had a cell size of < 5 observations in one or more cells of the 2x2 table, we would use a statistical test that is more valid under small sample sizes.

To perform a Fisher's Exact test for association:

```
fisher.test(data1$treat,data1$sex)
```

The R output is below:

```
        Fisher's Exact Test for Count Data

data:  dat$treat and dat$sex
p-value = 0.3787
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.687057 3.155144
sample estimates:
odds ratio
  1.457436
```

The p-value is 0.3787, which means that there is no statistically significant association between treatment group assigned and biological sex. Of course we expect these results to be very similar to the chi-square test given the sample size.

Chi-square tests for m x k tables

Now suppose we categorize age into one of three groups - < 40 years old, 40-64 years old, and 65+ years old. We can now look at the association between treatment group and age category.

Constructing the categorical Age variable:

```
data1$cat_age = as.factor(ifelse(data1$age < 40, '<40',
                          ifelse(data1$age < 65, '40-64',
                          ifelse(data1$age >= 65, '65+',  ' ' ))))
```

Construct the 2x3 table for age category and treatment group:

```
table_agextrt = table(data1$treat, data1$cat_age)
```

Which yields the following 2x3 table:

	<40	40-64	65+
Placebo	28	22	2
Treated	47	44	3

To perform a Chi-square test for association:

```
chisq.test(data1$treat,data1$cat_age)
```

The results are below:

```
        Pearson's Chi-squared test

data:  dat$treat and dat$cat_age

X-squared = 0.28834, df = 2, p-value = 0.8657

Warning message: In chisq.test(data1$treat, data1$cat_age) :
  Chi-squared approximation may be incorrect
```

Why did we get this warning message? As discussed previously, we know that chi-square tests are valid when there are at least 5 observations in each cell. In our 2x3 table, we see that two cells have counts < 5 (both in the age 65+ category as our patient population was relatively young). To remedy this, we will use a Fisher's Exact test, which is valid with small sample sizes.

```
fisher.test(data1$treat,data1$cat_age)
```

The R output is below:

```
        Fisher's Exact Test for Count Data
data:  data1$treat and data1$cat_age
p-value = 0.9051
alternative hypothesis: two.sided
```

Ordinal data

Because cat_age is ordinal in nature, we want to take advantage of this in our analysis when testing for an association. Why? Because if there is a true linear trend where increasing age groups have increasing probabilities to be in a particular treatment group, then this test is more powerful than a chi-square test, which just looks at overall association and does not take into account any natural ordering to the data.

To perform a Cochran-Armitage test for linear trend:

```
library("DescTools")
CochranArmitageTest(table_agextrt)
```

The results are below:

```
Cochran-Armitage test for trend

data:  table_agextrt

Z = -0.32743, dim = 3, p-value = 0.7433

alternative hypothesis: two.sided
```

McNemar's test for dependent categorical variables

Suppose now we have data on systolic blood pressure (BP) levels (categorized as normal or abnormal) that were collected on a randomly selected sample of individuals – pre and post exercise regimen.

Data:

	Post exercise	
Pre exercise	Normal	Abnormal
Normal	45	8
Abnormal	22	30

To enter this data in table format (*Note: you can always enter a table or the raw data for any of the tests we are discussing*) into R we code:

```
BPdata <- matrix(c(45, 22, 8, 30), nrow = 2,
dimnames = list("Pre" = c("Normal", "Abnormal"),
"Post" = c("Normal", "Abnormal")))
```

The goal is to examine whether exercise is associated with blood pressure status. As each individual contributes two observations to the analysis (pre and post BP status), we know that we cannot use a standard chi-square test for association, as this requires independent observations. Instead, we use McNemar's test, which takes the paired nature of the data into account.

To run a McNemar's test in R we use the following code:

```
mcnemar.test(BPdata)
```

The results are below:

```
        McNemar's Chi-squared test with continuity correction

data:  BPdata
McNemar's chi-squared = 5.6333, df = 1, p-value = 0.01762
```

The p-value is 0.018, therefore we observed a statistically significant association between exercise and blood pressure status.