Session 5: Statistical Analysis in R

Objectives:

By the end of this session, participants will be proficient in the following:

- Performing a chi-square test of association on categorical variables and interpreting results
- Carrying out a t-test on 2 groups of continuous variables and interpreting results
- Understanding confidence intervals
- Interpretation of p-values and how they relate to statistically significant results

Chi-sq test

The chi-square test is a test of association among two or more categorical variables

By performing a chi-square analysis, we are testing the null hypothesis that there is no association between chronic hepatitis C virus (HCV) infection and gender. The alternative is that there is.

Step 1: Set working directory

```
setwd("/Users/zainabsiddiq/Dropbox/Mac/Documents/Modeling workshop/data files")
```

Substitute your own PATH between quotation marks above

Step 2: Read in the dataset

```
hcv_data_clean <- read.csv("hepCdata_clean.csv")

ls()

dir()

head(hcv_data_clean)
```

Step 3: Explore the dataset

Look at the help file and arguments for a chi-squared test before we delve in further

```
help(chisq.test)

args(chisq.test)
```

Step 4: Creating a 2x2 as input for the chi-square function

Create an object using the table function to summarize Disease_Status and Gender

```
table.DiseaseSex <- table(hcv_data_clean$Disease_Status, hcv_data_clean$Sex)
```

View the table

```
table.DiseaseSex
```

Step 5: Running the Chi-square test

Now, we can use this table as an input to the chisq.test function to test whether Disease_Status and Gender are associated, using the chisq.test function.

```
chisq.result <- chisq.test(table.DiseaseSex)
```

The result is stored as chisq.result

Print the result

```
chisq.result
```

Exercise 5.1

Carry out a chi-square test to determine whether age is associated with Hepatitic C.

Use hcv_data_clean$AgeGroup and hcv_data_clean$Disease_Status

Store the result of the chi-square test in chisq2.result

T-TEST for Continuous variables

The two sample t-test is a comparison of the means of two groups. It is appropriate when sample sizes in each group are small. However, it does make some assumptions about the data being analyzed. The standard two sample t-test assumes that the data in each group are normally distributed and that their variances are similar.

We will now conduct a t-test analysis on our dataset to determine whether **the mean age of Hepatitis C patients differs from the mean age of Healthy controls**.

Step 1: Set working directory

```
setwd("/Users/zainabsiddiq/Dropbox/Mac/Documents/Modeling workshop/data files")
```

Substitute your own PATH between quotation marks above

Step 2 : Explore the dataset

Look at the help file and arguments for a t -test before we delve in further

```
help(t.test)

args(t.test)
```

Step 3 : Run the t-test

Run a t test to determine if mean Age varies by Disease_Status

```
ttest.results  <- with(hcv_data_clean, t.test(Age[Disease_Status == 'Healthy'], Age[Disease_Status == 'HepC']))
```

Show the t-test results

```
ttest.results
```

Exercise 5.2

AST is a liver enzyme, which is often elevated in people with chronic hepatitis C.

Run a t test to determine if AST levels varies by Disease_Status in our sample dataset.

Use variable names AST and Disease_Status for this analysis

Store the result of the t-test in ttestast.result