# Data Science and Statistics in Research: unlocking the power of your data

Session 2.3: How significant are your results?

## Introduction

In this session we will be creating confidence intervals and performing hypothesis tests on some example data.

## Preliminaries

We need the following package

- `Rmisc` - Package to create confidence intervals from samples.

Make sure that this package is downloaded and installed in `R`. We use the `require()` function to load it into the `R` library.

```
# Loading packages
require(Rmisc)
```

## Sampling Variability

Sample statistics can be used to estimate the characteristics of the underlying population or process from which the samples are drawn. We must be aware of two factors when attempting to make generalisations from a sample to a population of interest; bias and chance.

Suppose that we measure height (in centemetres) of the adult population of Bath, which is 100,000 residents. The true population mean height is 177cm and true standard deviation of 10. We simulate this data using the `rnorm()` function.

```
# Creating Simulated heights
heights <- rnorm(100000, # Population of one hundred thousand
                 mean = 177, # True mean height
                 sd=10) # True standard deviation
```

More information about the `rnorm()` function can be found by typing `?rnorm` into `R`.

Suppose we are studying heights of adults in Bath and take a sample of 50 people. We can take a sample of 50 people from the population of Bath using the `sample()` function.

```
# Taking a sample of the population
samp <- sample(heights, # Dataset of heights
               size=50) # Take 50 samples
```

We can calculate the sample mean, $\bar{x}$, which will be an estimate of the true population mean height in Bath. In turn the sample standard deviation, $s$, is an estimate of the true standard deviation of heights in Bath. We can calculate these statistics using the `mean()` and `sd()` functions respectively.

```
# Sample Mean
mean(samp)
[1] 179.4453
```

```
# Sample Standard Deviation
sd(samp)
[1] 12.16892
```

More information about the `sample()` function can be found by typing `?sample` into `R`.

This sample has sample mean 179.4 and sample standard deviation 12.2. We can see that the sample and the population means differ. Let's repeat this by taking another sample to see if we have similar results.

```
# Taking a sample of the population
samp <- sample(heights, # Dataset of heights
               size=50) # Take 50 samples
# Sample Mean
mean(samp)
[1] 176.5872

# Sample Standard Deviation
sd(samp)
[1] 8.986289
```

This sample has sample mean 176.6 and sample standard deviation 9. Again we can see that the sample and the population means differ.

We must be aware of two factors when attempting to make generalisations from these samples to population of Bath; bias and chance.

Bias is important in the planning of a study, where we must be clear about any possible selection effects that may bias a sample. If a sample is biased it will not representative of the population as a whole. For example, if our sample is from the basketball team where players are generally taller we cannot say this is representative of the population of Bath.

Chance is unavoidable. Where variability is present, the sample statistics calculated from any particular sample will be different to those calculated from another independent sample.

If we repeatedly take random samples from the overall population and each time record the mean and the standard deviation we would find that the sample mean and standard deviation vary from sample to sample, which is what we observe here.

## Confidence Intervals

When using samples of data, the sample mean will be the best estimate of the true population mean although there will be some uncertainty associated with this. A confidence interval quantifies this uncertainty and gives a range of values in which we are 'confident' the true population value will lie.

It is use to indicate the level of precision of an estimate from a sample, with larger samples giving more precision. A wider interval means that there is more uncertainty and often they are used to assess whether a particular value is likely or not.

To demonstrate how to construct confidence intervals in `R`, we will use the `mtcars` dataset. The `mtcars` dataset comprises of fuel consumption and other aspects of automobile design and performance for 32 cars recorded in the 1974 Motor Trend US magazine. We load this dataset, which is stored within `R`, using the `data()` function.

```
# Loading mtcar dataset
data(mtcars)
```

Make sure you are familiar with the contents of this dataset before continuing on with the rest of this session. Information on this dataset can be found by typing `?mtcars` into `R`.

Suppose a car company wants to estimate the mean fuel consumption (in miles per gallon) of the 32 cars tested. We can do this using the `mean()` function.

```
# Sample mean fuel consumption
mean(mtcars$mpg)
[1] 20.09062
```

This sample mean is the best estimate of the true population mean, but we want to construct a 95% confidence interval to give a range of values in which we am confident the true mean fuel consumption of cars in 1973-74 lie. To do this, we use the `CI()` function.

```
# Calculating a 95% confidence interval for fuel
# consumption in miles per gallon
CI(mtcars$mpg, # Data to create a confidence interval from
   0.95)       # Level of confidence interval
   upper     mean    lower
22.26357 20.09062 17.91768
```

More information about the `CI()` function can be found by typing `?CI` into `R`.

Using this ouput we are 95% 'confident' the true mean fuel consumption of cars in 1973-74 is between 17.92 and 22.26.

Note that the confidence interval for the fuel consumption is symmetric around the sample mean.

```
# Calculating the confidence interval
CI <- as.vector(CI(mtcars$mpg,0.95))

# Mean minus the lower confidence interval
CI[2] - CI[1]
[1] -2.172946

# Upper confidence interval minus the mean
CI[3] - CI[2]
[1] -2.172946
```

**Activities**

- Find 80%, 90% and 99% confidence intervals for the fuel consumption. You should see that the confidence intervals get wider as we increase the confidence. Why?
- Choose another continuous variable (for example weight) from `mtcars` datasets and adapt the above code to create confidence intervals.

# Hypothesis Testing

To test hypotheses in research, two alternative conclusions are set up, and on the basis of the experiment one is accepted and the other rejected. This acceptance and rejection is always on the basis of some pre-specified levels of confidence. There are 2 hypotheses constructed, the null and alternative.

The **null hypothesis**, $H_0$: we hypothesise that there is no difference between

- your sample and a population mean
- the mean of 2 groups.

The **alternative hypothesis**, $H_1$: we hypothesise that there is a difference between

- your sample and a population mean
- the mean of 2 groups.

The hypotheses are always stated so that either one or the other (but not both) can be true. On the basis of the hypotheses a statistical decision rule is constructed

- **IF** the observed value of a statistic takes on a certain range of values
- **THEN** we have enough evidence to reject the null hypothesis
- **OTHERWISE** there is not enough evidence to reject the null hypothesis.

There are two types of hypothesis test; one-tailed and two-tailed. One-tailed tests allow for the possibility of an effect in just one direction. For example, we would like to test whether true population mean is significantly **higher** or **lower** than a particular value. Two-tailed tests, you are testing for the possibility of an effect in two directions – both positive and negative.

## One-sample t-tests

A one sample t-test is used to determine whether the mean of a sample significantly differs from a known population mean.

To demonstrate how to perform a one-sample t-test in `R`, we will use the `mtcars` dataset. The `mtcars` dataset comprises of fuel consumption and 10 other aspects of automobile design and performance of 32 cars from 1973-74. We load this dataset, which is stored within `R`, using the `data()` function.

```
# Loading mtcar dataset
data(mtcars)
```

Make sure you are familiar with the contents of this dataset before continuing on with the rest of this session. Information on this datasets can be found by typing `?mtcars` into `R`.

Suppose a car company has tested 32 cars and thinks that the true mean fuel consumption (in miles per gallon) is 22.5. We use the `mean()` function to calculate the sample mean.

```
# Sample mean fuel consumption
mean(mtcars$mpg)
[1] 20.09062
```

Let's test whether there is a significant difference between the sample mean and the hypothesised mean. We construct the hypotheses

- **null**: the true mean fuel consumption in 1973-74 is 22.5
- **alternative**: the true mean fuel consumption in 1973-74 is not 22.5.

We choose a significance level of 0.05 and construct the statistical decision rule

- **IF** the p-value is less than 0.05
- **THEN** we have enough evidence to reject the null hypothesis
- **OTHERWISE** there is not enough evidence to reject the null hypothesis.

To perform a one-sample t-test, we use the `t.test()` function.

```
# Perform a one-sample t-test.
t.test(x  = mtcars$mpg, # Sample to be tested
       mu = 22.5, # Hypothesised mean
       significance = 'two.sided') # Option to test that the means are equal

    One Sample t-test

data:  mtcars$mpg
```

```
t = -2.2614, df = 31, p-value = 0.0309
alternative hypothesis: true mean is not equal to 22.5
95 percent confidence interval:
 17.91768 22.26357
sample estimates:
mean of x
 20.09062
```

Information on the `t.test()` function can be found by typing `?t.test` into `R`.

The p-value is less than 0.05, therefore there is enough evidence to reject the null hypothesis. We conclude the true population mean is not 22.5.

**Activities**

- Suppose we performed this test with significance of 0.1, 0.025 and 0.01. Would our conclusions have changed?
- To ensure you are comfortable with performing t-tests, perform a one-sample t-test but with different hypothesised means.

Suppose a car company has tested 32 cars and thinks that the true mean fuel consumption (in miles per gallon) is less than 22.5. We can reformulate the hypothesis used above to test for this

- **null**: the true mean fuel consumption in 1973-74 is 22.5
- **alternative**: the true mean fuel consumptionin 1973-74 is less than 22.5.

We choose a significance level of 0.05 and construct the statistical decision rule

- **IF** the p-value is less than 0.05
- **THEN** we have enough evidence to reject the null hypothesis
- **OTHERWISE** there is not enough evidence to reject the null hypothesis.

As before we perform a one-sample t-test, using the `t.test()` function but instead specify the option to say that we want to test whether the true mean is significantly less than the hypothesised mean.

```
# Perform a one-sample t-test.
t.test(x   = mtcars$mpg, # Sample to be tested
       mu = 22.5, # Hypothesised mean
       alternative  = 'less') # Specifying less than

    One Sample t-test

data:  mtcars$mpg
t = -2.2614, df = 31, p-value = 0.01545
alternative hypothesis: true mean is less than 22.5
95 percent confidence interval:
     -Inf 21.89707
sample estimates:
mean of x
 20.09062
```

More information on the `t.test()` function can be found by typing `?t.test` into `R`.

The p-value is less than 0.05, therefore there is enough evidence to reject the null hypothesis. We conclude the true population mean is less than 22.5.

**Activities**

- Suppose we performed this test with significance of 0.1, 0.025 and 0.01. Would our conclusions have changed?

- Perform a hypothesis test to test that the true fuel consumption (in miles per gallon) is greater than 22.5.