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

Session 2.5: Choosing the right tests for your data

Introduction

In this session we will be performing more hypothesis tests on some example data.

Preliminaries

We need the following package

- `ggplot2` - Package to implement the ggplot language for graphics in `R`.

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(ggplot2)
```

Independent t-test

An independent t-test to determine whether the means of the two independent samples significantly differ. Two samples are independent when one does not depend on the other.

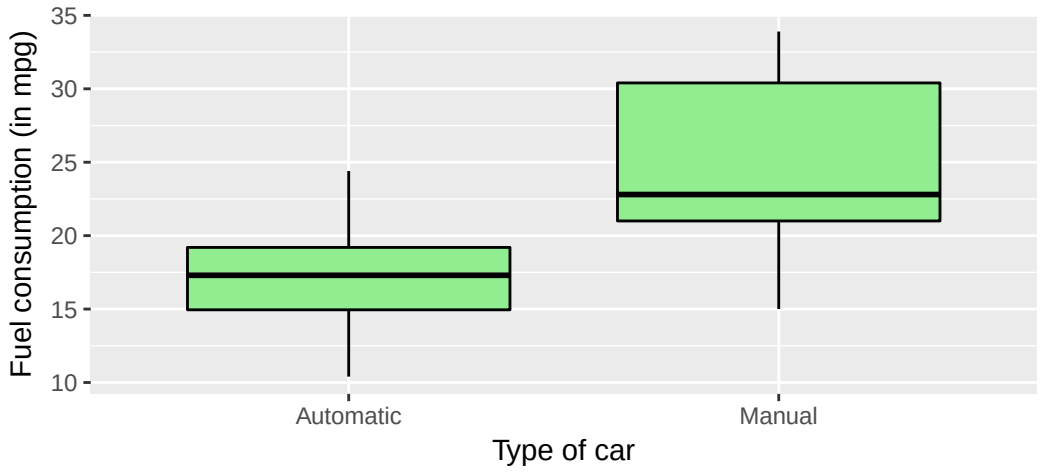
To demonstrate how to perform a independent t-test in `R`, we will use the `mtcars` dataset. The `mtcars` dataset comprises of fuel consumption and 10 other aspects of automobile design and performance of 32 cars from 1973-74. We load this dataset, which is stored within `R`, using the `data()` function.

```
# 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 wants to see if the true mean fuel consumption (in miles per gallon) differs significantly in automatic and manual cars. We can visualise the distribution of fuel consumption for both automatic and manual cars using boxplots. To do this we use the `ggplot2` package.

```
# Boxplot of fuel consumption using ggplot
ggplot(mtcars, aes(factor(am),mpg)) + # ggplot with the desired data
  geom_boxplot(fill='lightgreen',colour='black') + # Specifying boxplot
  labs(x="Type of car",y="Fuel consumption (in mpg)") + # Axes labels
  scale_x_discrete("Type of car", labels = c("Automatic","Manual"))
```



Information about the `ggplot2` package can be found in the 'Packages' pane.

Let's calculate the sample mean fuel consumption in automatic and manual cars separately. We do this by using the `aggregate()` function.

```
# Calculating the mean miles per gallon
# for automatic and manual cars separately
aggregate(mpg ~ am,    # Splitting mpg by automatic/manual
          data=mtcars, # Dataset to summarise
          FUN = mean)  # Summary statistic to use
  am      mpg
1  0 17.14737
2  1 24.39231
```

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

We observe that the distribution of fuel consumption differs between automatic and manual cars. Let's test whether the true means significant differ. We construct the hypotheses

- **null**: the true mean fuel consumption is the same for automatic and manual cars
- **alternative**: the true mean fuel consumption differs for automatic and manual cars

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.

The automatic cars sampled are not related to manual cars, so the samples are independent, therefore we use a independent t-test to test this hypothesis. We need to subset the data to separate the data into automatic and manual cars. To do this, we use the `subset()` function. We then use the `t.test()` function, to perform an independent t-test.

```
# Subsetting mtcars to extract all automatic cars
auto <- subset(mtcars, am == 0)

# Subsetting mtcars to extract all manual cars
man <- subset(mtcars, am == 1)

# Perform an independent t-test.
t.test(auto$mpg, # Fuel consumption for automatic cars
       man$mpg, # Fuel consumption for manual cars
       paired = FALSE, # Samples are independent not paired
```

```
       alternative = 'two.sided') # Say we want a two-tailed test

    Welch Two Sample t-test

data:  auto$mpg and man$mpg
t = -3.7671, df = 18.332, p-value = 0.001374
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.280194  -3.209684
sample estimates:
mean of x mean of y
 17.14737  24.39231
```

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

The p-value is less than 0.05, therefore have enough evidence to reject the null hypothesis. We conclude that the true mean fuel consumption is different between automatic and manual cars.

Activity

- Suppose we performed to test whether the true mean fuel consumption (in miles per gallon) of automatic cars is less than manual cars. Reformulate the hypothesis and alter the code above to perform the one-tailed independent t-test.

Paired t-test

A paired t-test is used to compare the means of two paired groups, when the two groups are of equal size and subjects in one sample are paired with one in the other.

To demonstrate how to perform a paired t-test in `R`, we take data from a study which was performed to assess the effect of different diets on LDL cholesterol in men. In total, 12 men went under two diets, with a 'washout' period in between them. After completion of the diets their LDL cholesterol level was measured.

```
# Cholesterol after diet 1
diet1<-c(4.61, 6.42, 5.40, 5.54, 3.98, 3.82, 5.01, 4.34, 3.80, 4.56, 5.35, 3.89)

# Cholesterol after diet 2
diet2<-c(3.84, 5.57, 5.85, 4.80, 3.68, 2.96, 4.41, 3.72, 3.49, 3.84, 5.26, 3.73)
```

Suppose a dietician wanted to test whether the true mean cholesterol differs after the two diets. Let's calculate the sample mean cholesterol level for both diets.

```
# Sample means of the two diets
mean(diet1); mean(diet2)
[1] 4.726667
[1] 4.2625

# Sample means of the two diets
mean(diet1 - diet2)
[1] 0.4641667
```

We can see that the sample means differ between the two diets. Let's test whether the population means differ significantly. We construct the hypotheses

- **null**: there is no difference in the true mean cholesterol between the diets
- **alternative**: there is a difference in the true mean cholesterol between the diets.

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 all the men in the study go on the same two diets, with their cholesterol measured after each, this data is paired. Therefore we should use a paired t-test to test this hypothesis. To do this, we use the `t.test()` function.

```
# Perform a paired t-test.
t.test(diet1, # Data from diet 1
       diet2, # Data from diet 2
       paired=TRUE, # Selecting that this is a paired test
       alternative="two.sided") # Selecting that we want a two-sided test

    Paired t-test

data:  diet1 and diet2
t = 4.0874, df = 11, p-value = 0.001798
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2142201 0.7141132
sample estimates:
mean of the differences
              0.4641667
```

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 have enough evidence to reject the null hypothesis. We conclude that the true mean level of cholesterol differs between the diets.

Activity

- Test the same hypothesis using an independent t-test rather than a paired t-test. Do you come to the same conclusion?

Chi-squared test

A Chi-Squared test is used to compare the proportions of outcomes in different groups. It tells you whether there is a difference or an association in the proportions between the groups being tested.

To demonstrate how to perform a paired t-test in `R`, we take data from a study that was performed to assess the effects of smokeless tobacco use and low-birth weight in India. We have frequencies of low-birth-weight babies given by categories of mothers' tobacco usage per week.

```
# Frequency Table
FreqTable <- rbind(c(646,160),c(85,27),c(35,21))

# Adding row names to the dataset
rownames(FreqTable) <- c('Non-users','1-4 times','5 or more times')

# Adding column names to the dataset
colnames(FreqTable) <- c('Regular Birth Weight','Low-birth Weight')

# Printing frequency table
FreqTable
```

```
                Regular Birth Weight Low-birth Weight
Non-users                        646              160
1-4 times                         85               27
5 or more times                   35               21
```

Suppose a doctor wanted to test whether there is an association between smokeless tobacco use and low-birth rate. Let's view this frequency table as proportions to see whether there are differences in the proportions of low-birth weight in each category of smokeless tobacco use. To do this, we use the `prop.table()`.

```
# Proportions
prop.table(FreqTable,1)
                Regular Birth Weight Low-birth Weight
Non-users                 0.8014888        0.1985112
1-4 times                 0.7589286        0.2410714
5 or more times           0.6250000        0.3750000
```

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

We observe that the proportions for low-birth rate increase as smokeless tobacco use increases. Let's test to see if this is significant. We construct the hypotheses

- **null**: there is no difference in the proportions between low-birth rate and smokeless tobacco use (i.e. the outcomes are independent and there is no association)
- **alternative**: alternative: there is a difference in the proportions between low-birth rate and smokeless tobacco use (i.e. the outcomes are dependent and there is an association).

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.

We use a Chi-Squared test to test this hypothesis. To do this, we use the `chisq.test()` function.

```
# Perform a Chi-Squared test.
chisq.test(FreqTable)

    Pearson's Chi-squared test

data:  FreqTable
X-squared = 10.282, df = 2, p-value = 0.005852
```

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

The p-value is less than 0.05, therefore reject the null hypothesis in favour of the alternative. We conclude that the proportions are not the same between the groups and there is an association between low-birth weight and smokeless tobacco use.