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

Session 2.7: When and how to use non-parametric statistics

Introduction

In this session we will be performing non-parametric tests on some example data.

Preliminaries

We need the following packages

- `ggplot2` - Package to implement the ggplot language for graphics in `R`
- `MASS` - Package for support functions and extra datasets.

Make sure that these packages are downloaded and installed in `R`. We use the `require()` function to load it into the `R` library.

```
# Loading packages
require(ggplot2)
require(MASS)
```

Non-parametric tests

The statistical tests you have seen so far require that the data can be assumed to follow a particular distribution, often the Normal distribution. This type of testing is called parametric.

There may be situations where we have data that is clearly non-Normal or it might be Normal, but there is not enough data to establish this. Parametric methods are usually fine to use with reasonably-sized samples as long as the data are unimodal and roughly symmetric about the mean. If data are severely non-Normal and sample sizes are small, these methods may be unreliable.

Non-parametric tests essentially compare medians rather than means, and use a rank order of observations. They make no assumptions about the underlying distributions of the data. They may also be suitable for comparing nominal (categorical) and ordinal (ordered categorical) data.

Checking for Normality

There are a few visual checks you can do to decide whether parametric and non-parametric tests are most appropriate.

We can use histograms and Q-Q Plots to look at the distribution of your data. If a histogram is symmetric around a value then you can consider your data normally distributed. If the dots on a Q-Q plot follow the straight line then your data is normally distributed.

We now look at two examples, one where normality is a good assumption for the sample and one where it is not.

Fuel consumption in `mtcars` dataset

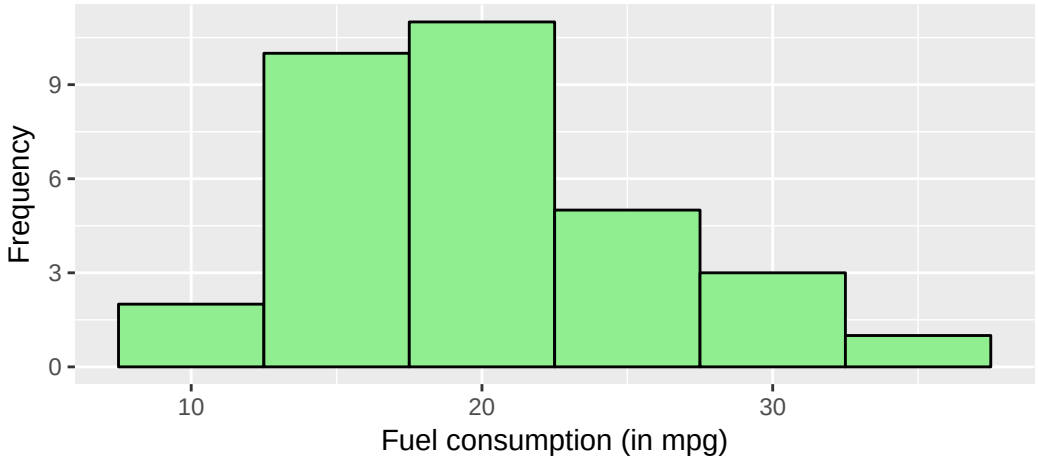
The `mtcars` dataset comprises of fuel consumption and 10 other aspects of automobile design and performance of 32 cars from 1973-74. We load these datasets, which are stored within `R`, using the `data()` function.

```
# Loading mtcar datasets
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`.

We are interested in whether normality is a good assumption in the fuel consumption (in miles per gallon) for 32 cars in `mtcars` datasets. We plot a histogram of the fuel consumption to check the distribution. To do this, we use the `ggplot2` package.

```
# Creating a histogram plot of fuel consumption in the mtcars dataset
ggplot(mtcars, aes(mpg)) + # Specify the data
  geom_histogram(binwidth = 5,
                 colour='black', fill='lightgreen') + # Specify we want a histogram
  labs(x='Fuel consumption (in mpg)',y='Frequency') # Axes Labels
```

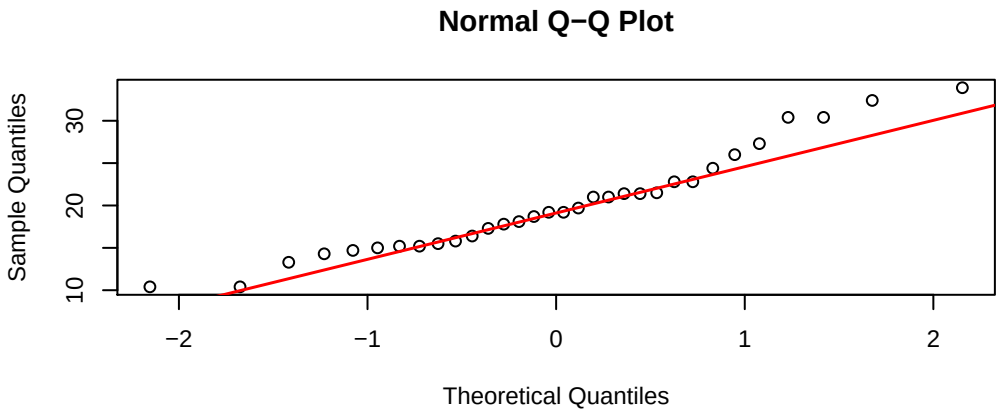


Information on the `ggplot2` package can be found in the 'Packages' pane.

We can see that the fuel consumption is roughly symmetric around the mean, so we can assume that it is Normally distributed.

We can also create a Q-Q plot to check the distribution further. To do this, we use the `qqnorm()` and `qqline()` functions.

```
# Creating a Q-Q plot of fuel consumption in the mtcars dataset
qqnorm(mtcars$mpg)
qqline(mtcars$mpg,
       col = 'red',
       lwd = 1.5)
```



More information on the `qqnorm()` and `qqline()` functions can be found by typing `?qqnorm` and `?qqline` into `R`.

We can see that the dots roughly follow the line. There are some deviations at the lower and higher values, but as the data is fairly symmetric and the points at the middle of the line follow the qqline, we can assume it is Normally distributed.

Prices in `diamonds` dataset

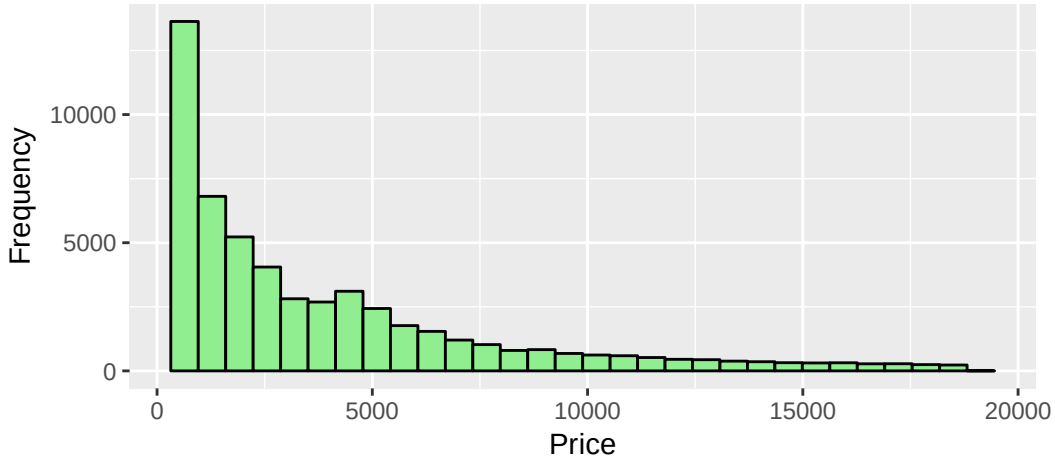
The `diamonds` dataset contains prices and other attribures for 53,940 diamonds. We load these datasets, which are stored within `R`, using the `data()` function.

```
# Loading diamonds datasets
data(diamonds)
```

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

We are interested in whether normality is a good assumption in the prices of diamonds in the `diamonds` datasets. We plot a histogram of the prices to check the distribution. To do this, we use the `ggplot2` package.

```
# Creating a histogram plot of price in the diamonds dataset
ggplot(diamonds, aes(price)) + # Specify the data
  geom_histogram(colour='black', fill='lightgreen') + # Specify we want a histogram
  labs(x='Price',y='Frequency') # Axes Labels
```

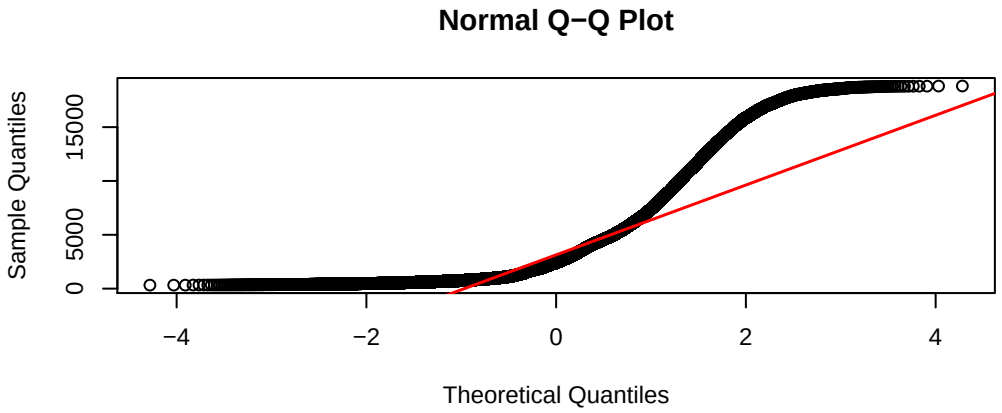


Information on the `ggplot2` package can be found in the 'Packages' pane.

We can see the distributions are heavily skewed, and can't be negative, so we conclude that Normality cannot be assumed here.

We can further study the Normality of the distribution using a qqplot.To do this, we use the `qqnorm()` and `qqline()` functions.

```
# Create a Q-Q plot of prices in the diamonds dataset
qqnorm(diamonds$price)
qqline(diamonds$price,
       col = 'red',
       lwd = 1.5)
```



More information on the `qqnorm()` and `qqline()` functions can be found by typing `?qqnorm` and `?qqline` into `R`.

We can see that the points of the data severely deviate from the qqline. Because of this, we cannot assume Normality in this case.

One-sample Wilcoxon Signed Rank test

A one-sample Wilcoxon Signed Rank Test is used to determine whether the median of a sample significantly differs from a specified value when there is evidence of non-normality.

To demonstrate how to perform a one-sample Wilcoxon Signed Rank test in `R`, we will use the `immer` dataset. The `immer` dataset comprises of yields of five varieties of barley that were grown in six locations in 1931 and 1932. We load this dataset, which is stored within `R`, using the `data()` function.

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

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 `?immer` into `R`.

A farmer grows five varieties of barley in six locations in 1931 and 1932 and measures the yearly yield. The farmer thinks the true median yield of barley in 1931 is 117. We use the `median()` function to calculate the sample median barley yield in 1931.

```
# Calculating the median miles per gallon
median(immer$Y1)
[1] 102.95
```

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

- **null**: the true median barley yield in 1931 is 117
- **alternative**: the true median barley yield in 1931 is not 117.

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

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

There are not enough observations to deduce whether the data is normally distributed so we use a non-parametric test. As we are testing for the true median using one sample, we use a one-sample Wilcoxon Signed Rank test. To do this, we use the `wilcox.test()` function.

```
# One sample Wilcox Rank test
wilcox.test(immer$Y1,
            mu=117,
            alternative = 'two.sided')

    Wilcoxon signed rank test

data:  immer$Y1
V = 145, p-value = 0.07324
alternative hypothesis: true location is not equal to 117
```

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

As the p-value is less than 0.05 we do not have enough evidence to reject the null hypothesis and conclude that there is no evidence to show the true median barley yield is not 117 in 1931.

Activities

- Suppose we performed this test with significance of 0.1, 0.025 and 0.01. Would our conclusions have changed?
- Suppose we want to test that the true median barley yield in 1931 is less than 117. Reformulate the hypothesis and alter the code above to perform the one-tailed one sample Wilcoxon Signed Rank Test. Would our conclusions have changed?

Mann Whitney U Test

A two-sample Mann Whitney U test is used to compare the distribution of a numeric variable between two groups when there is evidence of non-Normality.

To demonstrate how to perform a two-sample Wilcoxon Signed Rank test in `R`, we will use the `diamonds` dataset. The `diamonds` dataset comprises of prices and other attributes of 53,940 diamonds. We load this dataset, which is stored within `R`, using the `data()` function.

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

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 `?diamonds` into `R`.

A jeweller sells diamonds and records the price that they sold for in addition to attributes of the diamond such as a cut grading which are either 'Fair' 'Good', 'Very Good', 'Premium' or 'Ideal'. The jeweller wants to test whether the median prices of diamonds with a 'Fair' cut are different from those with a 'Good' cut.

Let's calculate the sample median prices of diamonds with a 'Fair' cut are different from those with a 'Good' cut. We first need to subset the data to extract 'Fair' cut diamonds and 'Good' cut diamonds from the

data. To do this, we use the `subset()` function. We then use the `median()` function, to calculate the sample medians.

```
# Extracting all the 'Fair' cut diamonds
dat1 <- subset(diamonds, cut == 'Fair')

# Extracting all the 'Good' cut diamonds
dat2 <- subset(diamonds, cut == 'Good')

# Median price for 'Fair' cut diamonds
median(dat1$price)
[1] 3282

# Median price for 'Good' cut diamonds
median(dat2$price)
[1] 3050.5
```

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

We observe that the sample median of diamonds with a 'Fair' cut are different from those with a 'Good' cut. Let's test whether this is significant. We construct the hypotheses

- **null**: the true median price between diamonds with a 'Fair' cut is the same from those with a 'Good' cut
- **alternative**: the true median price between diamonds with a 'Fair' cut is different from those with a 'Good' cut

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 prices are severely skewed so we use non-parametric tests. The sampled diamonds with a 'Fair' cut are not related to those with a 'Good' cut, so the samples are independent, therefore we use a Mann Whitney U test to test this hypothesis. To do this, we use the `wilcox.test()` function.

```
# Mann Whitney U Test
wilcox.test(dat1$price, # Prices of `Fair' cut diamonds
            dat2$price, # Prices of `Good' cut diamonds
            paired = FALSE, # Data is not paired
            alternative = 'two.sided') # Say we want a two-tailed test

    Wilcoxon rank sum test with continuity correction

data:  dat1$price and dat2$price
W = 4441700, p-value = 5.551e-14
alternative hypothesis: true location shift is not equal to 0
```

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

The p-value is less than 0.05 we have enough evidence to reject the null hypothesis and conclude that that the true median price between diamonds with a `Fair ' cut are different from those with a`Good' cut.

Activities

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

- Suppose we want to test that diamonds with a 'Fair' cut are less than those with a 'Good' cut. Reformulate the hypothesis and alter the code above to perform the one-tailed one-sample Wilcoxon Signed Rank Test.

Two-sample Wilcoxon Signed Rank test

A two-sample Wilcoxon Signed Rank Test is used to compare the distribution of a numeric variable of two groups when there is evidence of non-Normality. The two groups must be of equal size and subjects in one sample are paired with one in the other.

To demonstrate how to perform a two-sample Wilcoxon Signed Rank test in `R`, we will use the `immer` dataset. The `immer` dataset comprises of yields of five varieties of barley that were grown in six locations in 1931 and 1932. We load this dataset, which is stored within `R`, using the `data()` function.

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

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 `?immer` into `R`.

A farmer grows five varieties of barley in six locations in 1931 and 1932 and measures the yearly yield and wants to test if the amount yields in 1931 and 1932 are different.

- **null**: there is no difference in the true median yield between 1931 and 1932
- **alternative**: there is a difference in the true median yield between 1931 and 1932.

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.

There are not enough observations to deduce whether the data is normally distributed so we use a non-parametric test. The yields are of the same five varieties of barley grown in the same six locations in 1931 and 1932, this data is paired. Therefore we should use a two-sample Wilcoxon Signed Rank test to test this hypothesis. To do this, we use the `wilcox.test()` function.

```
# Two sample Wilcox Rank test
t.test(immer$Y1, # Yield from 1931
            immer$Y2, # Yield from 1932
            paired = TRUE, # Data is not paired
            alternative = 'two.sided') # Say we want a two-tailed test

    Paired t-test

data:  immer$Y1 and immer$Y2
t = 3.324, df = 29, p-value = 0.002413
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  6.121954 25.704713
sample estimates:
mean of the differences
               15.91333
```

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

The p-value is less than 0.05 we have enough evidence to reject the null hypothesis and conclude that that there is a difference in the true median yield between 1931 and 1932.

Activities

- Test the same hypothesis using an independent t-test rather than a Mann Whitney U-test. Do you come to the same conclusion?
- Suppose we want to test that the yield in 1931 is less than the yield in 1932. Reformulate the hypothesis and alter the code above to perform the one-tailed two-sample Wilcoxon Signed Rank Test.