National College of Ireland
Statistics for Data Analytics

Tutorial - One Way Analysis of Variance

Noel Cosgrave *Associate Faculty, National College of Ireland*

Problem

A researcher wants to know if there is any difference in the weekly total hours of study for students from three universities. The results of a random sample of students from each university are as shown in the table below. At a significance level of 0.05, can we conclude that there is evidence of a difference in the average study time between the three universities?

University A	University B	University C
6.4	4.2	8.3
6.5	5.1	8.1
5.2	6.0	8.0
7.5	4.4	5.2
4.9	5.0	6.7
6.0	6.9	7.2
5.6	7.0	7.7
5.9	6.8	7.1
	5.5	8.3
	5.1	

Hypotheses and test selection

We first establish the null and alternative hypotheses:

$H_0 : \mu_A = \mu_B = \mu_C$ - there is no difference in the average study time for students at the three universities

$H_1 : \exists i, j : \mu_i \neq \mu_j$ where $i, j \in \{ A, B, C \}$ - the average study time is different for at least one of the three universities.

As there are more than two independent samples, the most appropriate statistical test is the one-way ANOVA. If the assumptions of this test are not met, the non-parametric Kruskal-Wallis H test can be used instead.

Solution using R

Preparation

Most of the required functionality is provided by base R. We will use the *ggplot2* library to produce the plots for assessing normality and homogeneity of variance. However, it is worth noting that the base R provides functions that can produce simpler versions of these plots. We will also use the *car* library for the Levene test of homogeneity of variance. If *ggplot2* and *car* were not previously installed, they will need to be installed using the *install.packages()* function.

```
install.packages(c("car","ggplot2"))
```

One the libraries are installed, we can load them.

```
library (ggplot2)
library (car)
```

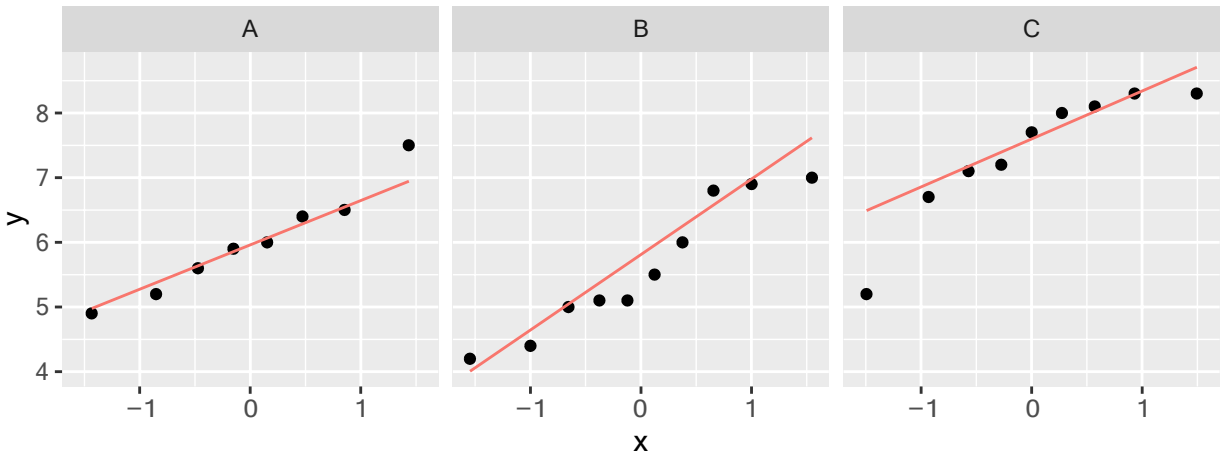
We also need to create variables to contain the data. In R, this is done using a long format data frame. In this format, there are two variables, one containing the group names and one for the values. Note that the *rep()* function repeats the supplied value a given number of times.

```
uni_df <- data.frame(
    University = c(rep("A",8),rep("B",10),rep("C",9)),
    Study =  c(6.4,6.5,5.2,7.5,4.9,6.0,5.6,5.9,
               4.2,5.1,6.0,4.4,5.0,6.9,7.0,6.8,5.5,5.1,
               8.3,8.1,8.0,5.2,6.7,7.2,7.7,7.1,8.3)
)
```

Check for normality

A fundamental assumption of one-way ANOVA is that the data for each group are drawn from a normally distributed population. We will first assess this using visualisations. Although multiple graphical methods can potentially be used to assess normality, the best is the Normal Quantile-Quantile plot.

```
ggplot(uni_df, aes(sample=Study)) +
  stat_qq() +
  stat_qq_line(aes(color="red")) +
  facet_wrap( ~ University, nrow = 1) +
  theme(legend.position = "none")
```



From these plots, we can see that the data points are dotted around the line, but diverge somewhat in the tails. Visual methods of assessing normality lack precision, leading to the possibility of incorrect conclusions.

Statistical methods of assessing normality are more reliable. We can check if the sample for each university is drawn from a normally-distributed population using the Shapiro-Wilk test. The hypotheses of this test are:

H_0: - The populations from which the sample is drawn does not differ significantly from a normal distribution.

H_1: - The populations from which the sample is drawn differs significantly from a normal distribution.

```
for (university in c("A","B","C")) {
    print(shapiro.test(uni_df[uni_df$University    ==    university,"Study"]))
}
```

```
    Shapiro-Wilk normality test

data:  uni_df[uni_df$University == university, "Study"]
W = 0.9695, p-value = 0.8941


    Shapiro-Wilk normality test

data:  uni_df[uni_df$University == university, "Study"]
W = 0.90868, p-value = 0.2721


    Shapiro-Wilk normality test

data:  uni_df[uni_df$University == university, "Study"]
W = 0.85458, p-value = 0.08368
```

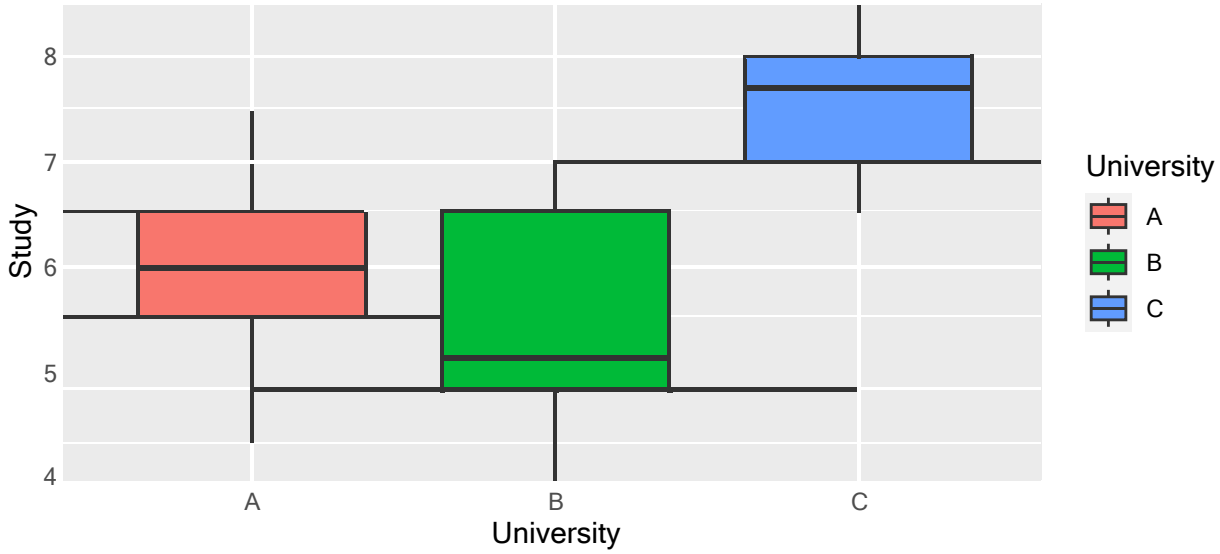
As the p value for each university is not significant, we fail to reject the null hypothesis and conclude that there is insufficient evidence to suggest that the samples are drawn from populations that are not normally distributed.

Check for homogeneity of variance

If the variances of the populations from which the samples were drawn are equal, we can use the regular one-way ANOVA. If there variances are significantly different, a one-way ANOVA with Welch's correction is required.

We can assess the variances using visual methods. The box and whisker plot is the most appropriate approach here.

```
ggplot(uni_df, aes(x=University, y=Study, fill=University)) +
    geom_boxplot()
```



We can see from this plot that the variance for University B is considerably larger than those for the other universities. However, just as was the case when visually assessing normality, conclusions drawn from such plots can be incorrect.

We can evaluate equality of variance using statistical methods.

Bartlett's test

This test is the most suitable for cases where there is strong evidence that the samples have been drawn from a normally-distributed population. It has the following hypotheses:

H_0: The variances of the populations from which the samples are drawn are the same.

H_1: The variance of at least one of the populations from which the samples are drawn is different.

```
bartlett.test(Study ~ University, data=uni_df)
```

```
    Bartlett test of homogeneity of variances

data:  Study by University
Bartlett's K-squared = 0.41972, df = 2, p-value = 0.8107
```

At 0.8107, the p value is greater than α at 0.05. We can conclude that there is no significant evidence of a difference between the variances of the populations from which the samples were drawn.

Levene's test

Levene's test is less sensitive to departures from normality than Bartlett's test, but is less powerful. Its hypotheses are:

H_0: The variances of the populations from which the samples are drawn are the same.

H_1: The variance of at least one of the populations from the samples are drawn is different.

```
leveneTest(Study ~ University, data=uni_df)
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  0.3306 0.7217
      24
```

At 0.7217, the p value is greater than α at 0.05. We can conclude that there is no significant evidence of a difference between the variances of the populations from which the samples were drawn.

Carry out a one way ANOVA

As we have established that the assumption of homogeneity of variance holds, we can use the standard one-way ANOVA. In R this is performed using the *aov()* function. We also produce a summary of the model using the *summary()* function.

```
res.aov <- aov(Study ~ University, data = uni_df)
summary(res.aov)
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
University    2  16.50   8.250   8.911 0.00128 **
Residuals    24  22.22   0.926
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is shown in the *Pr(>F)* column. At 0.00128, this is lower than α at 0.05. As such we can reject the null hypothesis and conclude that the average study time is different for at least one of the universities. As ANOVA is an omnibus test, we cannot yet tell where this difference lies. In order to establish this, we perform a *post-hoc* test.

Perform a post-hoc analysis

As the sample sizes are fairly similar, and given that the assumptions of normally and homogeneity of variance hold, we can use the Tukey Honest Significant Difference (HSD) test.

```
TukeyHSD(res.aov)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Study ~ University, data = uni_df)

$University
     diff         lwr       upr       p adj
B-A -0.4 -1.5397934 0.7397934 0.6600266
C-A  1.4  0.2324027 2.5675973 0.0166854
C-B  1.8  0.6959457 2.9040543 0.0012342
```

Here we examine the values in the *p adj* column. We can see that the p value for universities A and B is not significant, so we conclude that this is not where the difference lies. However, the p value for universities A and C as well as that for B and C are both significant, indicating that the average study times for these universities are likely to be different.

Solution using Python

Preparation

```
import scipy.stats as st
import matplotlib.pyplot as plt
```

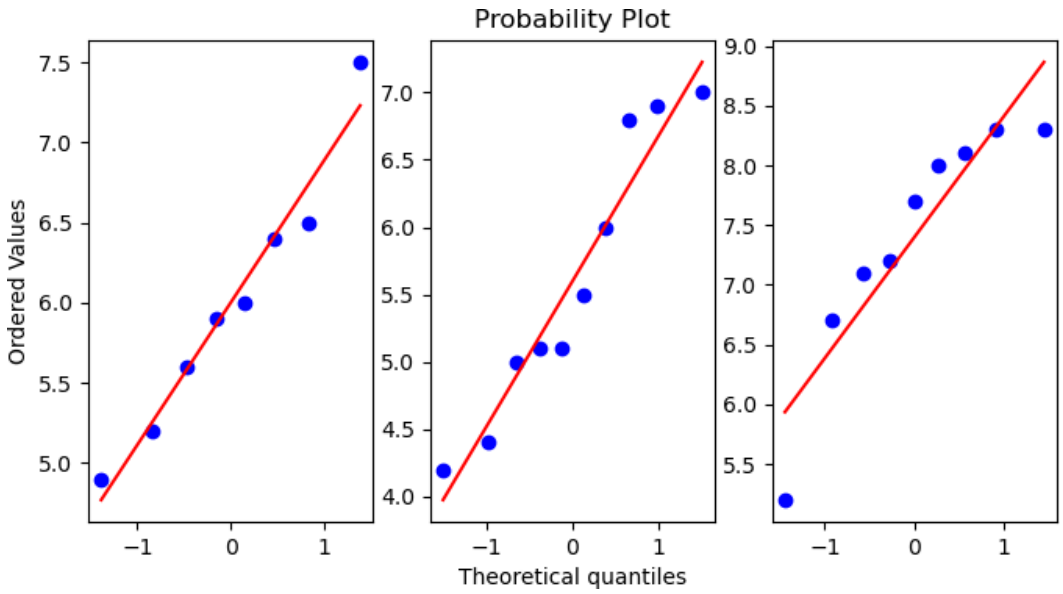
We also need to create variables to contain the data. In Python we create these as lists, with one list per sample. We can also create a list of lists containing the data for all universities.

```
university_A = [6.4,6.5,5.2,7.5,4.9,6.0,5.6,5.9]
university_B = [4.2,5.1,6.0,4.4,5.0,6.9,7.0,6.8,5.5,5.1]
university_C = [8.3,8.1,8.0,5.2,6.7,7.2,7.7,7.1,8.3]
universities  =  [university_A,  university_B,  university_C]
```

Check for normality

A fundamental assumption of one-way ANOVA is that the data for each group are drawn from a normally distributed population. We will first assess this using visualisations. Although multiple graphical methods can potentially be used to assess normality, the best is the Normal Quantile-Quantile plot, just as we did in the R example.

```
fig, (ax1,ax2,ax3) = plt.subplots(nrows=1, ncols=3, figsize=(8,4))
_ = st.probplot(university_A, dist="norm", plot=ax1)
_ = st.probplot(university_B, dist="norm", plot=ax2)
_ = st.probplot(university_C, dist="norm", plot=ax3)
_ = ax1.set(xlabel=None, title=None)
_ = ax2.set(ylabel=None)
_ = ax3.set(xlabel=None, ylabel=None, title=None)
plt.show()
```



As we saw earlier, the data points are dotted around the line, but diverge somewhat in the tails. In order to obtain a clearer picture of the normality assumption for our samples, we will use the Shapiro-Wilk test. The hypotheses of this test are:

H_0: - The populations from which the sample is drawn does not differ significantly from a normal distribution.

H_1: - The populations from which the sample is drawn differs significantly from a normal distribution.

The code below outputs the results of the Shapiro Wilk test for each university in turn.

```
for university in universities:
    st.shapiro(university)
```

```
ShapiroResult(statistic=0.9695030450820923, pvalue=0.8940800428390503)
ShapiroResult(statistic=0.9086835980415344, pvalue=0.2720703184604645)
ShapiroResult(statistic=0.8545814752578735, pvalue=0.08367527276277542)
```

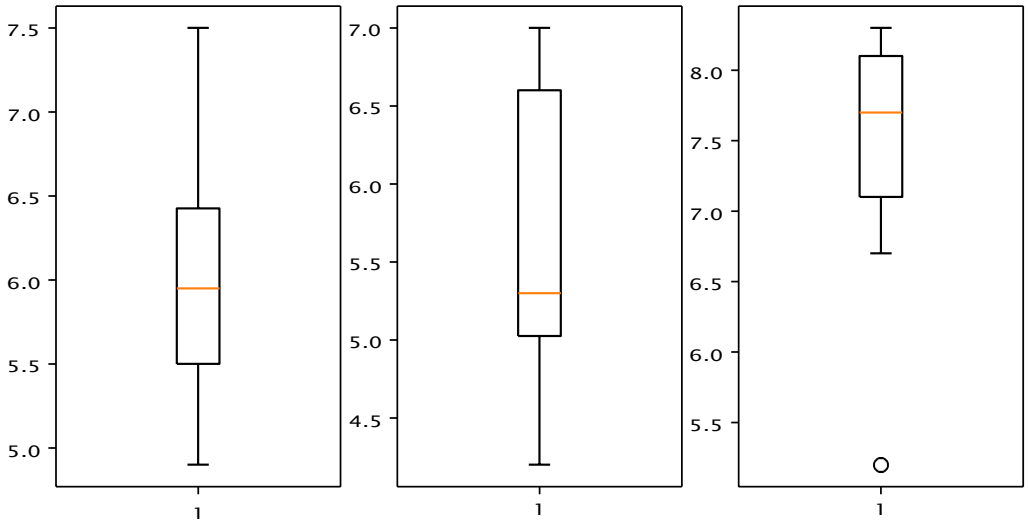
As the p value for each university is not significant, we fail to reject the null hypothesis and conclude that there is insufficient evidence to suggest that the samples are drawn from populations that are not normally distributed.

Check for homogeneity of variance

If the variances of the populations from which the samples were drawn are equal, we can use the regular one-way ANOVA. If there variances are significantly different, a one-way ANOVA with Welch's correction is required.

We can assess the variances using visual methods. The box and whisker plot is the most appropriate approach here.

```
fig, (ax1,ax2,ax3) = plt.subplots(nrows=1, ncols=3, figsize=(8,4))
_ = ax1.boxplot(university_A)
_ = ax2.boxplot(university_B)
_ = ax3.boxplot(university_C)
plt.show()
```



We can see from this plot that the variance for University B is considerably larger than those for the other universities. However, just as was the case when visually assessing normality, conclusions drawn from such plots can be incorrect.

We can evaluate equality of variance using statistical methods.

Bartlett's test

This test is the most suitable for cases where there is strong evidence that the samples have been drawn from a normally-distributed population. It has the following hypotheses:

H_0: The variances of the populations from which the samples are drawn are the same.

H_1: The variance of at least one of the populations from which the samples are drawn is different.

```
st.bartlett(university_A,university_B,university_C)
```

```
BartlettResult(statistic=0.41971650776537855,  pvalue=0.8106991512833053)
```

At approximately 0.8107, the p value is greater than α at 0.05. We can conclude that there is no significant evidence of a difference between the variances of the populations from which the samples were drawn.

Levene's test

Levene's test is less sensitive to departures from normality than Bartlett's test, but is less powerful. Its hypotheses are:

H_0: The variances of the populations from which the samples are drawn are the same.

H_1: The variance of at least one of the populations from the samples are drawn is different.

```
st.levene(university_A, university_B, university_C)
```

```
LeveneResult(statistic=0.33058324192856076, pvalue=0.7217246187778119)
```

At approximately 0.7217, the p value is greater than α at 0.05. We can conclude that there is no significant evidence of a difference between the variances of the populations from which the samples were drawn.

Carry out a one way ANOVA

As we have established that the assumption of homogeneity of variance holds, we can use the standard one-way ANOVA. In Python this is performed using the *f_oneway()* function from the *scipy stats* module.

```
st.f_oneway(university_A, university_B, university_C)
```

```
F_onewayResult(statistic=8.911291129112913,  pvalue=0.0012752777198787601)
```

At approximately 0.00128, the p value is lower than α at 0.05. As such we can reject the null hypothesis and conclude that the average study time is different for at least one of the universities. As ANOVA is an omnibus test, we cannot yet tell where this difference lies. In order to establish this, we perform a *post-hoc* test.

Perform a post-hoc analysis

As the sample sizes are fairly similar, and given that the assumptions of normally and homogeneity of variance hold, we can use the Tukey Honest Significant Difference (HSD) test.

```
res = st.tukey_hsd(university_A, university_B, university_C)
print(res)
```

```
Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval)
Comparison  Statistic  p-value  Lower CI  Upper CI
 (0 - 1)      0.400     0.660    -0.740     1.540
 (0 - 2)     -1.400     0.017    -2.568    -0.232
 (1 - 0)     -0.400     0.660    -1.540     0.740
 (1 - 2)     -1.800     0.001    -2.904    -0.696
 (2 - 0)      1.400     0.017     0.232     2.568
 (2 - 1)      1.800     0.001     0.696     2.904
```

The pairs are numbered according to the order in which they were entered into the model. (0-1) indicates universities A and B. We can see that the p value for this pair is not significant, so we conclude that this is not where the difference lies.

For the pair (0-2) or universities A and C, the p value is significant, indicating that the average study times for these universities are likely to be different.

For the pair (1-2) or universities B and C, the p value is significant, indicating that the average study times for these universities are likely to be different.