MSMS - 106

Ananda Biswas

Practical 08

✎ **Question :** A manufacturing company has purchased three new machines of different makes and wishes to determine whether one of them is faster than the others in producing a certain output. Five-hourly production figures are observed at random from each machine and the results are as follows.

	Machine A_1	Machine A_2	Machine A_3
Observations	25	31	24
	30	39	30
	36	38	28
	38	42	25
	31	35	28

Use analysis of variance technique and determine whether the machines are significantly different in their mean speeds. Use $\alpha = 5\%$.

One-way ANOVA

Key assumptions of ANOVA are that the errors (and consequently the observations) must be normally distributed with homoscedastic variance.

First we shall verify the assumptions.

```
machine_data <- read.csv('https://raw.githubusercontent.com/sakunisgithub/data_sets/refs/heads/master/msc_semester_1/sonam_madam_practical_08_data.csv',
                         stringsAsFactors = TRUE)
```

- Checking Normality

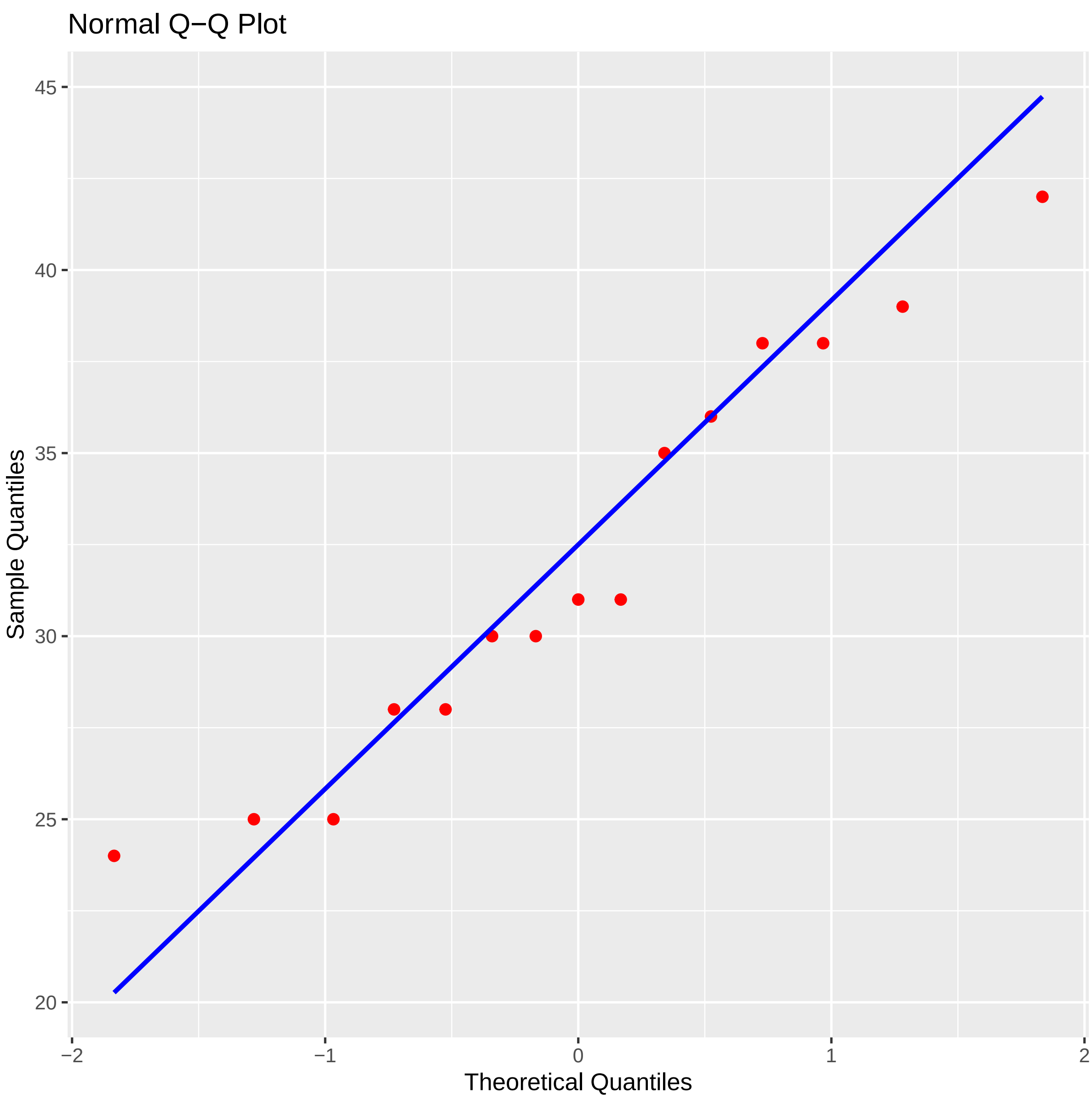
```
shapiro.test(machine_data$speed)

##
##  Shapiro-Wilk normality test
##
## data:  machine_data$speed
## W = 0.94173, p-value = 0.4046
```

A $p-$value of 0.4045684 results in failure of rejecting H_0 at 5% level of significance that the data is from a normal distribution. This can also be verified by a normal Q-Q plot.

```r
library(tidyverse)
```

```r
machine_data %>%
  ggplot(aes(sample = speed)) +
  stat_qq(size = 2, col = "red") +
  stat_qq_line(linewidth = 1, col = "blue") +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles", title = "Normal Q-Q Plot")
```



The points fit the line good.

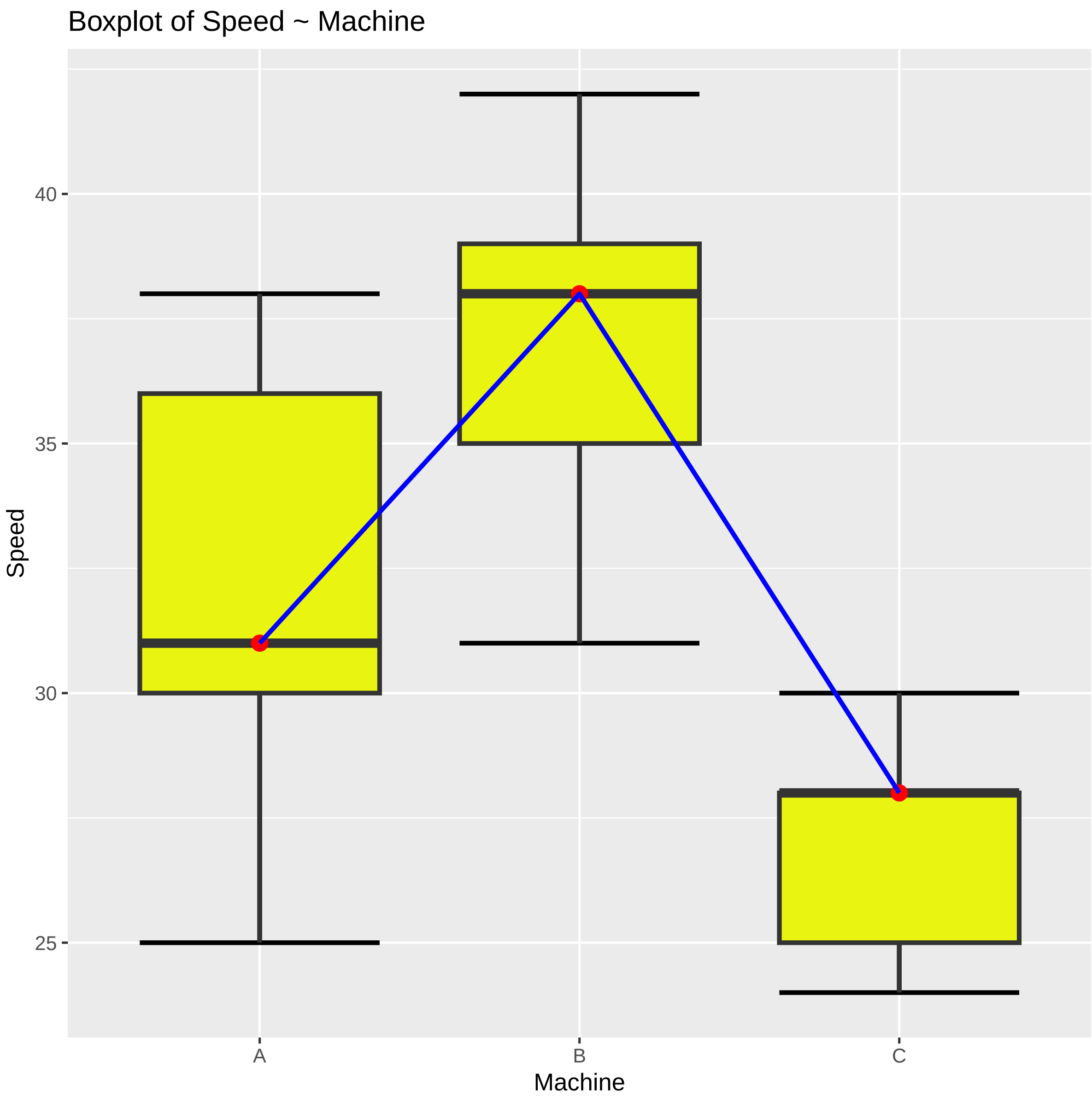
- Checking Homoscedastic Variance

```r
bartlett.test(speed ~ machine, data = machine_data)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  speed by machine
## Bartlett's K-squared = 1.8329, df = 2, p-value = 0.3999
```

A $p-$value of 0.3999448 results in failure of rejecting H_0 at 5% level of significance that all the group variances are equal. This can also be verified by a box-plot.

```
machine_data %>%
  ggplot(aes(x = machine, y = speed)) +
  stat_boxplot(geom = "errorbar", linewidth = 1) +
  geom_boxplot(fill = "#eaf411", linewidth = 1) +
  stat_summary(fun = median, geom = "point", size = 3, col = "red") +
  stat_summary(fun = median, geom = "line", aes(group = 1), linewidth = 1, col = "blue") +
  labs(x = "Machine", y = "Speed", title = "Boxplot of Speed ~ Machine")
```



The spread of the boxes are more or less similar across groups. The line joining the group averages (here median) clearly shows that the sample machine average speeds differ a lot.

```r
machine_data_anova <- aov(speed ~ machine, data = machine_data)
summary(machine_data_anova)

##             Df Sum Sq Mean Sq F value  Pr(>F)
## machine      2    250  125.00     7.5 0.00771 **
## Residuals   12    200   16.67
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$p-$value corresponding to machine is $0.0077073 < 0.05$. So we reject the null hypothesis of equality of mean machine speeds at 5% level of significance and conclude that at least one of the machine has significantly different mean speed than others.

Now we shall do pairwise comparisons.

```r
TukeyHSD(machine_data_anova, ordered = TRUE)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##     factor levels have been ordered
##
## Fit: aov(formula = speed ~ machine, data = machine_data)
##
## $machine
##     diff       lwr      upr     p adj
## A-C    5 -1.888394 11.88839 0.1709498
## B-C   10  3.111606 16.88839 0.0058028
## B-A    5 -1.888394 11.88839 0.1709498
```

$p-$value corresponding to comparison of Machine B and Machine C is $0.0058028 < 0.025$. So machine B and C are significantly different at 5% level of significance. Other comparisons are not significant. So we conclude that machine B is the best as it has the highest sample mean speed.