One-way ANOVA - Fixed Effects Model

Ananda Biswas

```
life_hours <- read.csv("D:\\data_sets\\life_hours_of_bulbs_data.csv",
    stringsAsFactors = TRUE)
```

```
life_hours
```

```
##    batch life_of_bulb
## 1      A         1600
## 2      A         1610
## 3      A         1650
## 4      A         1680
## 5      A         1700
## 6      A         1720
## 7      A         1800
## 8      B         1580
## 9      B         1640
## 10     B         1640
## 11     B         1700
## 12     B         1750
## 13     C         1460
## 14     C         1550
## 15     C         1600
## 16     C         1620
## 17     C         1640
## 18     C         1660
## 19     C         1740
## 20     C         1820
## 21     D         1510
## 22     D         1520
## 23     D         1530
## 24     D         1570
## 25     D         1600
## 26     D         1680
```

```
dim(life_hours)
```

```
## [1] 26  2
```

```
names(life_hours)
```

```
## [1] "batch"        "life_of_bulb"
```

```r
head(life_hours)
```

```
##   batch life_of_bulb
## 1     A          1600
## 2     A          1610
## 3     A          1650
## 4     A          1680
## 5     A          1700
## 6     A          1720
```

```r
tail(life_hours)
```

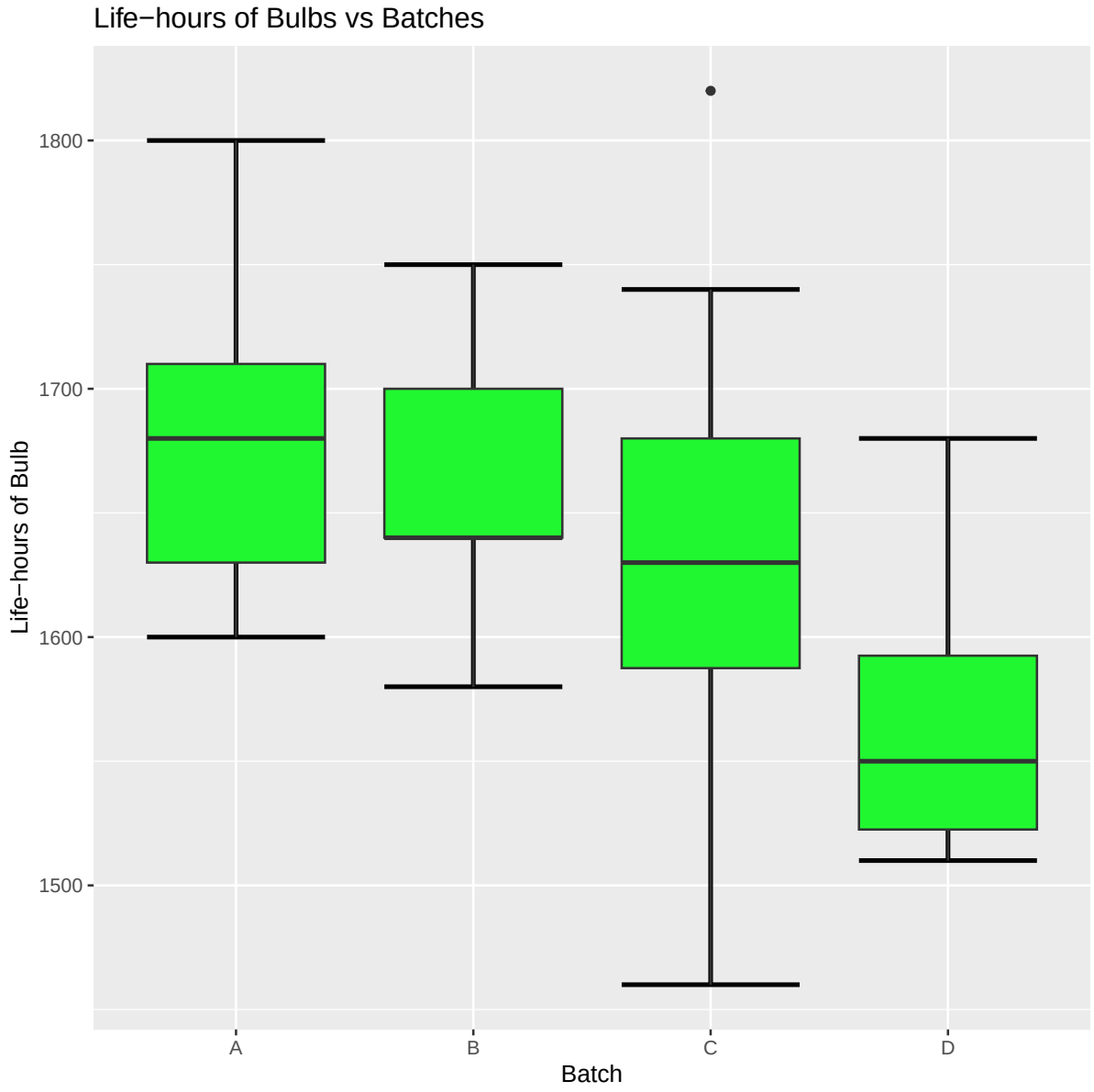
```
##    batch life_of_bulb
## 21     D          1510
## 22     D          1520
## 23     D          1530
## 24     D          1570
## 25     D          1600
## 26     D          1680
```

```r
summary(life_hours)
```

```
##  batch  life_of_bulb
##  A:7   Min.   :1460
##  B:5   1st Qu.:1585
##  C:8   Median :1640
##  D:6   Mean   :1637
##        3rd Qu.:1695
##        Max.   :1820
```

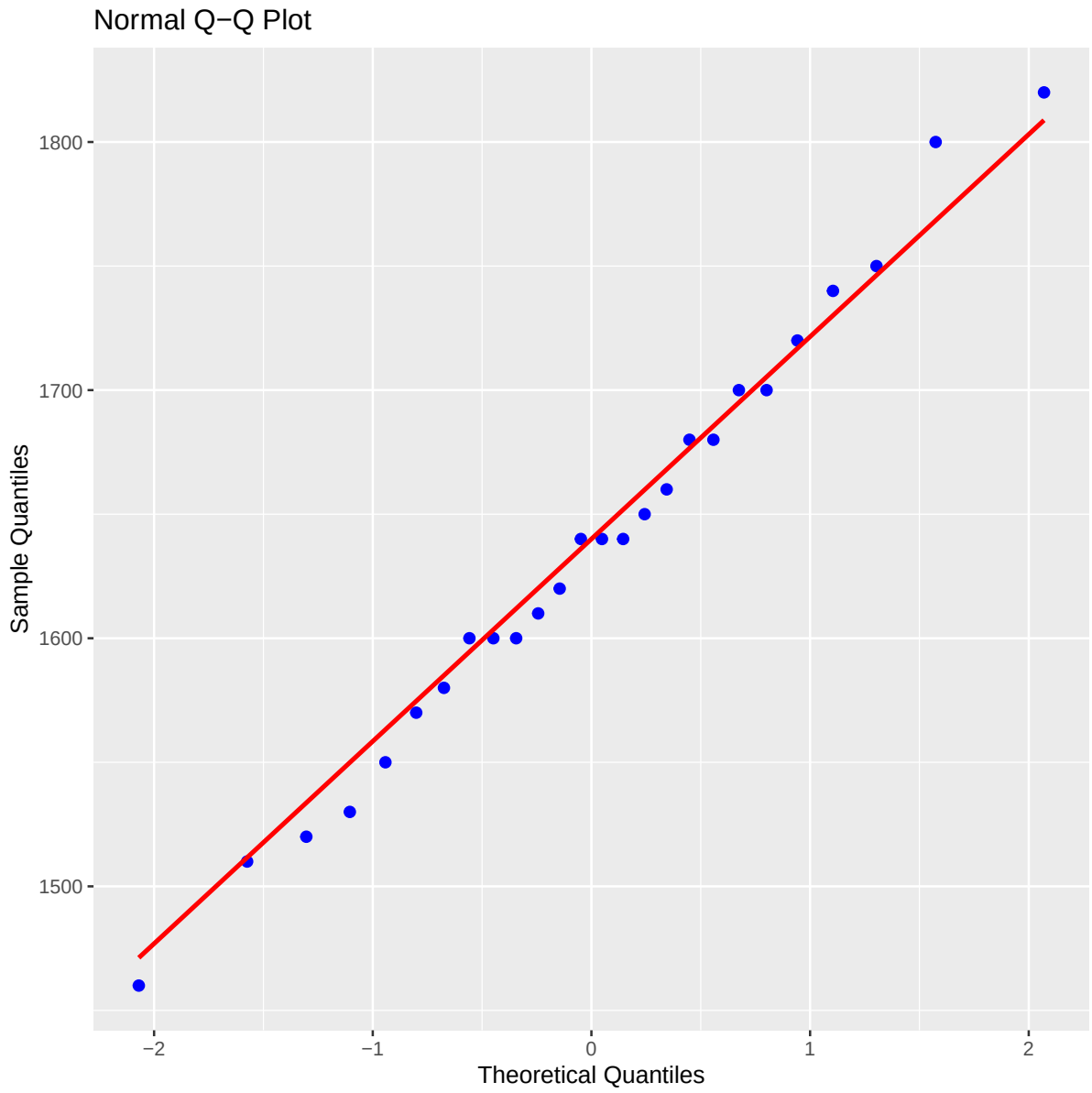
```
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.3
## Warning: package 'ggplot2' was built under R version 4.2.2
## Warning: package 'tibble' was built under R version 4.2.3
## Warning: package 'tidyr' was built under R version 4.2.3
## Warning: package 'readr' was built under R version 4.2.2
## Warning: package 'purrr' was built under R version 4.2.3
## Warning: package 'dplyr' was built under R version 4.2.3
## Warning: package 'stringr' was built under R version 4.2.3
## Warning: package 'forcats' was built under R version 4.2.2
## Warning: package 'lubridate' was built under R version 4.2.2
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0
--
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.1     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts()
--
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
life_hours %>%
    ggplot(aes(x = batch, y = life_of_bulb)) + stat_boxplot(geom = "errorbar",
    linewidth = 1) + geom_boxplot(fill = "#21F731") +
    labs(x = "Batch", y = "Life-hours of Bulb", title = "Life-hours of Bulbs vs Batches")
```



- **Testing Normality of Our Sample**

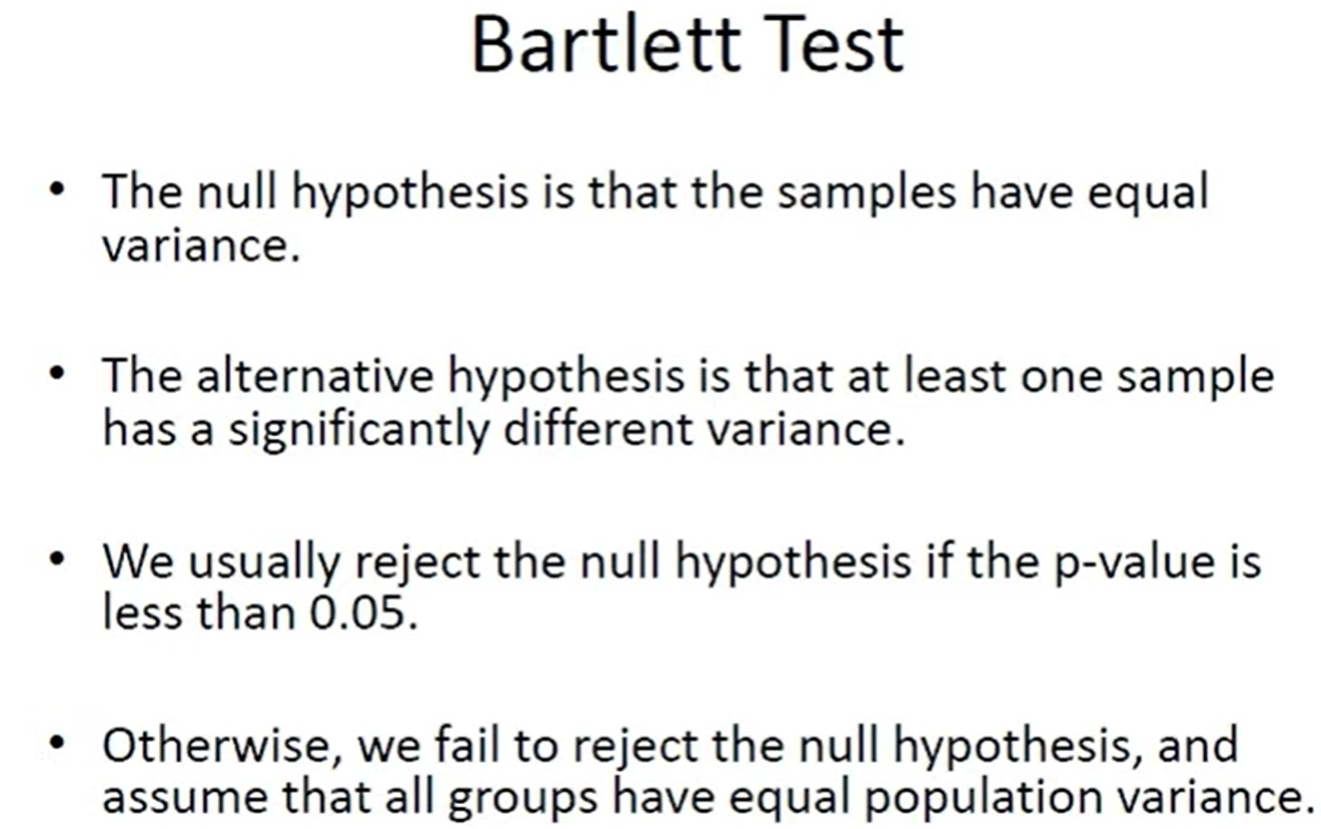
```
life_hours %>%
    ggplot(aes(sample = life_of_bulb)) + geom_qq(size = 2,
    col = "blue") + geom_qq_line(col = "red", linewidth = 1) +
    labs(x = "Theoretical Quantiles", y = "Sample Quantiles",
        title = "Normal Q-Q Plot")
```



We see that the line is a good fit. Hence, we conclude the sample is from a normal population.

- **Testing Equality of Several Population Variances (Homogenity)**

```
homo_test <- bartlett.test(life_of_bulb ~ batch, data = life_hours)
homo_test

##
##  Bartlett test of homogeneity of variances
##
## data:  life_of_bulb by batch
## Bartlett's K-squared = 2.508, df = 3, p-value = 0.4738
```




```
homo_test$p.value

## [1] 0.4738465
```

See that the p-value is much higher than $\alpha = 0.05$, so we fail to reject the null hypothesis and conclude that our homoscedastic assumption is true.

```r
fit1 <- lm(life_of_bulb ~ batch, data = life_hours)
summary(fit1)
```

```
##
## Call:
## lm(formula = life_of_bulb ~ batch, data = life_hours)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -176.250  -45.833   -8.125   36.417  183.750
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1680.00      31.35  53.589   <2e-16 ***
## batchB        -18.00      48.57  -0.371   0.7145
## batchC        -43.75      42.93  -1.019   0.3192
## batchD       -111.67      46.15  -2.420   0.0242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.94 on 22 degrees of freedom
## Multiple R-squared:  0.2267,Adjusted R-squared:  0.1212
## F-statistic: 2.149 on 3 and 22 DF,  p-value: 0.1229
```

```r
model.matrix(fit1)
```

```
##    (Intercept) batchB batchC batchD
## 1            1      0      0      0
## 2            1      0      0      0
## 3            1      0      0      0
## 4            1      0      0      0
## 5            1      0      0      0
## 6            1      0      0      0
## 7            1      0      0      0
## 8            1      1      0      0
## 9            1      1      0      0
## 10           1      1      0      0
## 11           1      1      0      0
## 12           1      1      0      0
## 13           1      0      1      0
## 14           1      0      1      0
## 15           1      0      1      0
## 16           1      0      1      0
## 17           1      0      1      0
## 18           1      0      1      0
## 19           1      0      1      0
## 20           1      0      1      0
## 21           1      0      0      1
## 22           1      0      0      1
## 23           1      0      0      1
```

```
## 24           1      0      0      1
## 25           1      0      0      1
## 26           1      0      0      1
## attr(,"assign")
## [1] 0 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$batch
## [1] "contr.treatment"
```

```
fit1$rank

## [1] 4
```

As the rank of the model matrix is 4, only 4 parameters have been estimated. *batchA* or α_1 has been forced 0.

```r
life_hours_anova <- aov(life_of_bulb ~ batch, data = life_hours)
summary(life_hours_anova)

##             Df Sum Sq Mean Sq F value Pr(>F)
## batch        3  44361   14787   2.149  0.123
## Residuals   22 151351    6880
```

See that, p-value corresponding to the test of equality of batch means is 0.123 which is much higher than $\alpha = 0.05$. So, we conclude that there is no significant difference between the batch means.

- **Pairwise Comparison**(although not necessary here)

```
TukeyHSD(life_hours_anova)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = life_of_bulb ~ batch, data = life_hours)
##
## $batch
##           diff       lwr       upr     p adj
## B-A  -18.00000 -152.8615 116.86146 0.9821643
## C-A  -43.75000 -162.9518  75.45182 0.7401446
## D-A -111.66667 -239.8048  16.47143 0.1025335
## C-B  -25.75000 -157.0525 105.55248 0.9470311
## D-B  -93.66667 -233.1322  45.79889 0.2714523
## D-C  -67.91667 -192.3036  56.47024 0.4452307
```

See that, all the p-values are greater than 0.05, implying that no two of the batch means differ significantly.

```r
df1 <- data.frame(batch = life_hours$batch, residuals = fit1$residuals)
```

```r
df1 %>%
    ggplot(aes(x = batch, y = residuals)) + geom_hline(yintercept = 0,
    col = "#FB2209", linewidth = 1) + stat_boxplot(geom = "errorbar",
    linewidth = 1) + geom_boxplot(fill = "#F10BCB") +
    labs(x = "Sample", y = "Residuals", title = "Residual Plot")
```