02441 Applied Statistics and Statistical Software

Exercise 3A - Kali

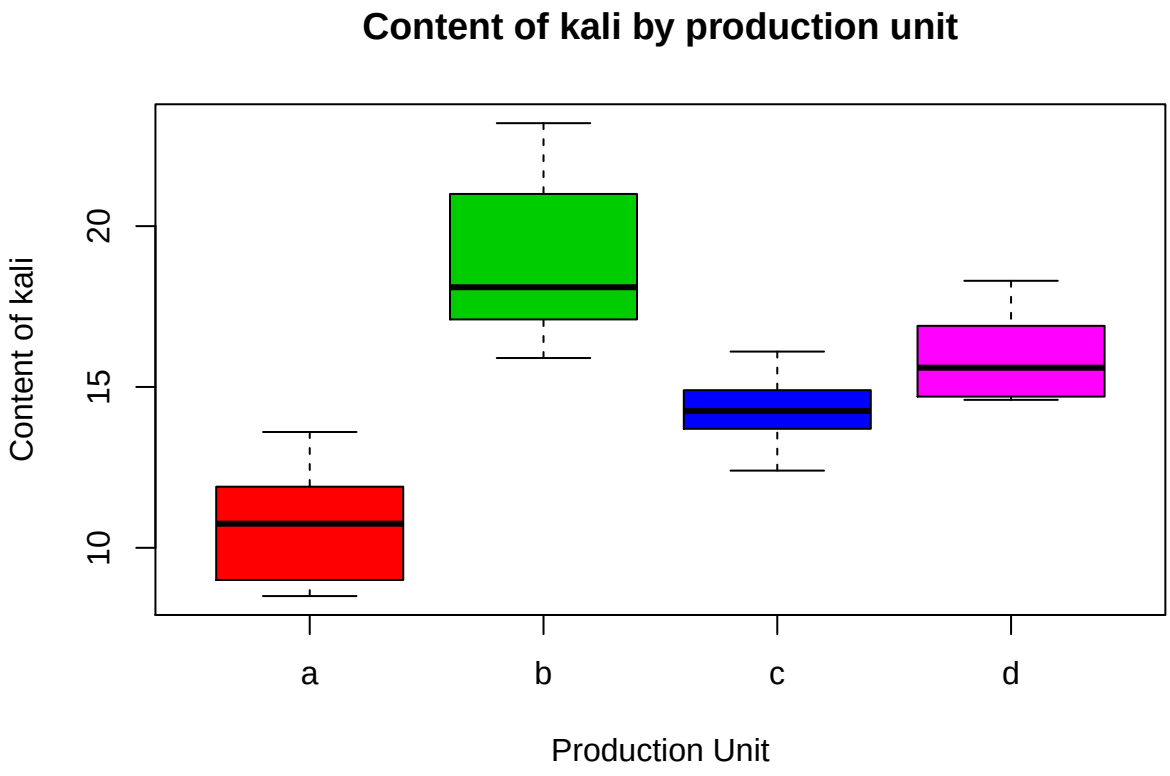
The dataset kali contains measurements of the content of kali (K2O) for four different productions

Variable name	Description
production	production unit
kali	content of kali

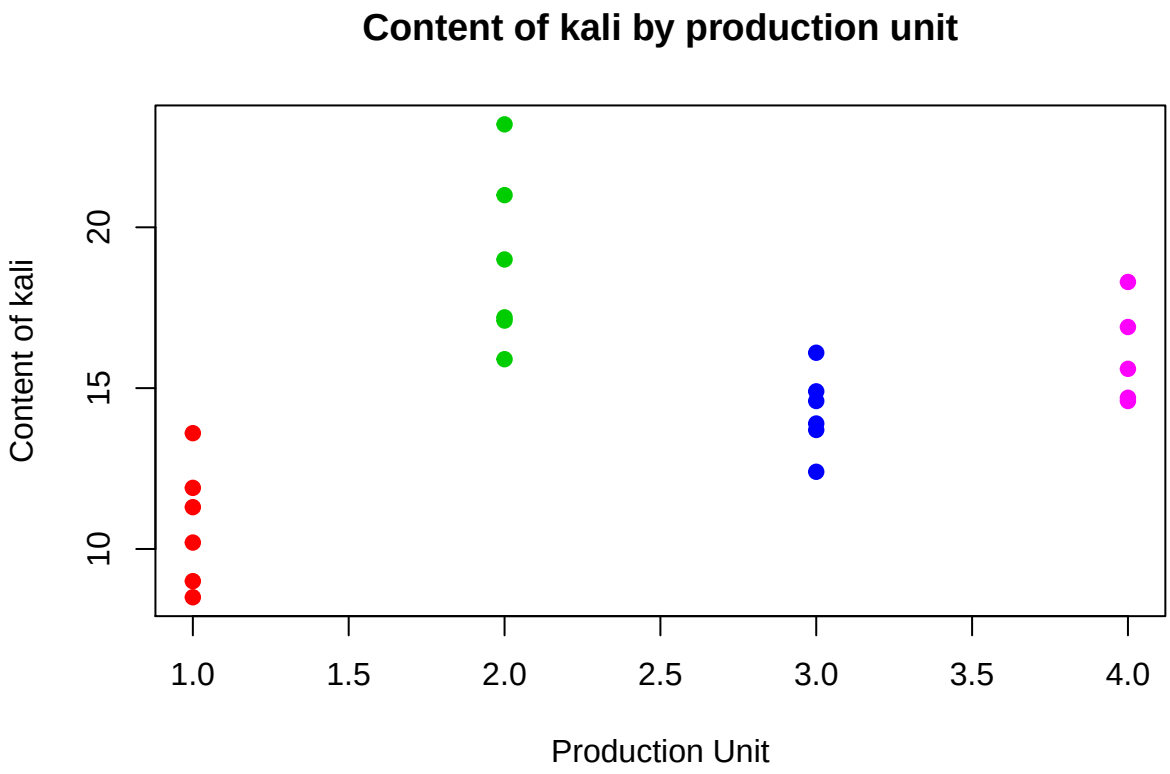
1. Use a non-parametric test to examine if the content of kali depends on the different productions

Start by loading and plotting the data

```
df <- read.table("kali.txt", header=TRUE)
kali <- df$kali
production <- df$production
boxplot(kali ~ production, xlab = "Production Unit", ylab = "Content of kali", col = c(2,3,4,6), main="
```




```
plot(as.numeric(production), kali, col = c(rep(c(2,3,4,6), each = 6)), xlab="Production Unit", ylab="Co
```



Non-parametric test

```
kruskal.test(kali~production)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  kali by production
## Kruskal-Wallis chi-squared = 17.853, df = 3, p-value = 0.0004717
```

Distributions are the same since p-value is less than $\alpha = 0.05$, i.e. we cannot reject H_0.

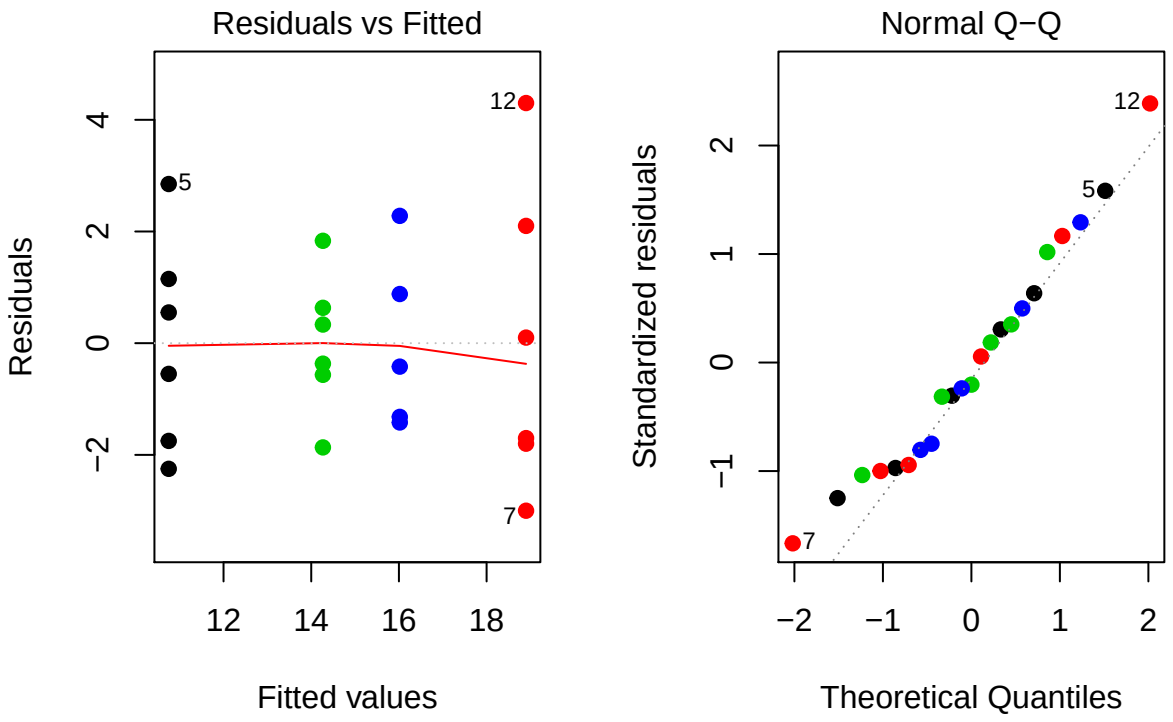
2. Use a one-way ANOVA to examine if the content of kali depends on the different productions

```
# Calculate one-way ANOVA
lm1 <- lm(kali~production, df)
anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: kali
##            Df  Sum Sq Mean Sq F value    Pr(>F)
```

```
## production  3 207.978  69.326  17.815 9.485e-06 ***
## Residuals  19  73.936   3.891
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Checking model assumptions
par(mfrow = c(1,2))
plot(lm1, col = df$production, which = 1:2, pch=19)
```



One-way ANOVA shows that there's statistical significance of the content of kali depending on the production of different productions; the means of the different groups are the same, i.e. p > α=0.05. The graph also shows that there's similar variance of both groups and their residuals are normally distributed.

3. If the content of kali depends on the different productions, which of the production(s) yield the highest content?

```
summary(lm1)
```

```
##
## Call:
## lm(formula = kali ~ production, data = df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.0000 -1.5600 -0.3667  1.0150  4.3000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.7500     0.8053  13.348 4.20e-11 ***
## productionb   8.1500     1.1389   7.156 8.42e-07 ***
## productionc   3.5167     1.1389   3.088  0.00606 **
## productiond   5.2700     1.1945   4.412  0.00030 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.973 on 19 degrees of freedom
## Multiple R-squared:  0.7377, Adjusted R-squared:  0.6963
## F-statistic: 17.82 on 3 and 19 DF,  p-value: 9.485e-06
```

Production b, since the intercept is the sum of production a + production b, i.e. $10.75 + 8.15 = 18.9$.