Dallin Romney – u1087199
Design of Experiments – ME EN 5035
February 26, 2019

Homework 4 – Blocking and Linear Regression

Part 1. Open Statistics 3rd Edition Problems:

6.41 – Open Source Textbook:

a) Hypotheses:
   a. Null: There is no significant correlation between the professor's prediction and the actual results
   b. Alternative: There is a significant correlation between the professor's prediction and the actual results
b) He predicted that 0.6*126 = 75.6 (76) buy the hard copy, 0.25*126 = 31.5 (31) print it, and 0.15*126 = 18.9 (19) read it online.
c) Assumptions:
   a. Data are non-timed, singularly counted frequencies/counts ☑
   b. Levels are mutually exclusive (e.g., no students said they did two of the options) ☑
   c. Independent study groups ☑
   d. 2 variables, with categories ☑ (expected vs actual)
   e. 5 or more values in at least 80% of the categories, and none with less than 1 expected ☑
d) Test:
   a. Chi-squared statistic: 1.0046
   b. Degrees of Freedom: (2-1)*(3-1) = 2 DOF
   c. p-value: 0.605124 → not significantly different!
e) With a confidence threshold of 95%, the professor's predictions and the actual textbook statistics are not significantly different. Thus, he made a pretty good prediction.

6.48 – Coffee and Depression:

a) Chi square test
b) Hypotheses:
   a. Null: There is no significant correlation between the coffee consumption and depression
   b. Alternative: There is a significant correlation between coffee consumption and depression
c) 0.05138 (5.14%) suffer from clinical depression, while 0.94863 (94.86%) do not.
d) Observed = 373, expected = 6617/50739*2607 = 340.0 → $(373 - 340)^2/340 = 3.203$
e) With (2-1)(5-1) = 4 degrees of freedom, a chi square statistic of 20.93 corresponds to a p-value of 0.000327
f) With a confidence threshold of 95%, there is a significant difference between the proportions of depressed women based on how much coffee they drink – coffee drinkers are less depressed.
g) I agree with this statement. The variables might not be independent – perhaps people drink less coffee *because* they are depressed, not vice versa (there could be other correlation not causation considerations, too!). Also, coffee is a food with a potentially addictive chemical – encouraging people to drink it based off of such a study may be short-sighted. Tons of factors can cause depression, and many factors can be caused by drinking coffee.

7.17 – Correlation, Part I:

a) If men always married women 3 years younger than themselves, the correlation would be linear and positive, with a slope of 1 and an intercept of -3.
b) If men always married women 2 years older than themselves, the correlation would be linear and positive, with a slope of 1 and an intercept of 2.
c) If men always married women half as old as themselves, the correlation would be linear and positive, with a slope of 0.5 and an intercept of 0.

7.30 – Cats, Part I:

a) Linear model: Heart Weight [g] = 4.034 [g/kg]*(Body weight [kg]) – 0.357 g
b) Intercept: The weight of a heart if there was no body (meaningless) – how much the model is raised from intercepting the origin
c) Slope: the increase of heart weight in grams correlated with an increase of 1 kg in body weight
d) R^2: The model explains 64.66% (64.41% adjusted) of the variability of the data around the linear model
e) R = sqrt(R^2) = 0.8041 or 80.41%

7.36 Beer and Blood Alcohol Content:

a) There is a positive linear correlation between BAC and cans of beer consumed
b) Linear model: BAC [g/deciliter] = 0.018 [g/deciliter/can]*(Cans consumed) – 0.0127 g/deciliter
    a. Intercept: The BAC if no cans of beer were consumed (probably not correct, should be 0)
    b. Slope: the increase of BAC in g/deciliter correlated with drinking a can of beer
c) Test:
   Hypotheses:
    a. Null: there is not a significant correlation between BAC and cans of beer consumed
    b. Alternative: there is a significant correlation between BAC and cans of beer consumed

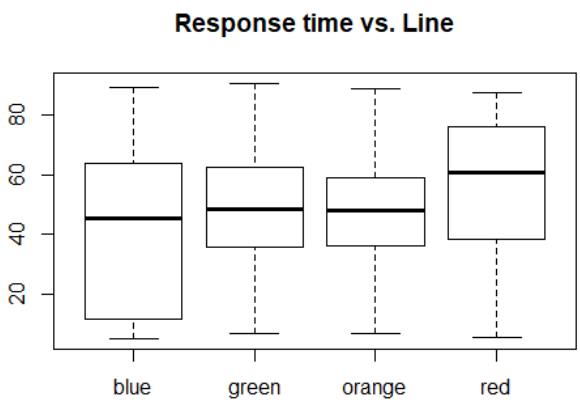
   p-value ~= 0 → There is a significant correlation between BAC and cans of beer consumed
d) R = 0.89 → R^2 = 0.7921 or 79.21% of the variability (residual) is explained by the linear model
e) There is no way of knowing – it could be stronger or weaker. There would be no way to verify if they were telling the truth, and they may be consuming drinks with a different level of alcohol than the beer in the study. Also, they may be older than the college students, which could affect BAC.

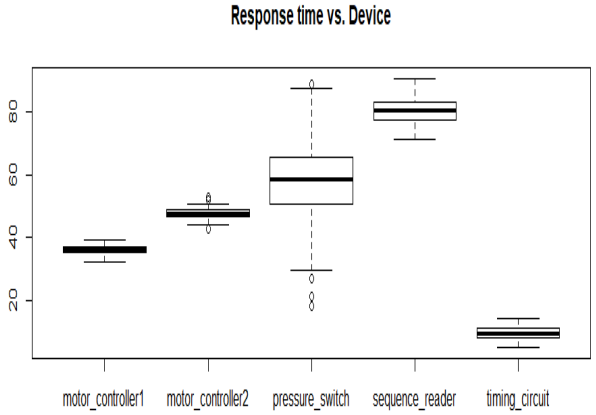
Part 2. Assembly Line Fault Investigation

Introduction

To analyze the effect of assembly line and device type on response time, data is visualized ANOVA and pairwise tests are performed on the linear models, and assumptions are verified. See attached code.

Data visualization





ANOVA

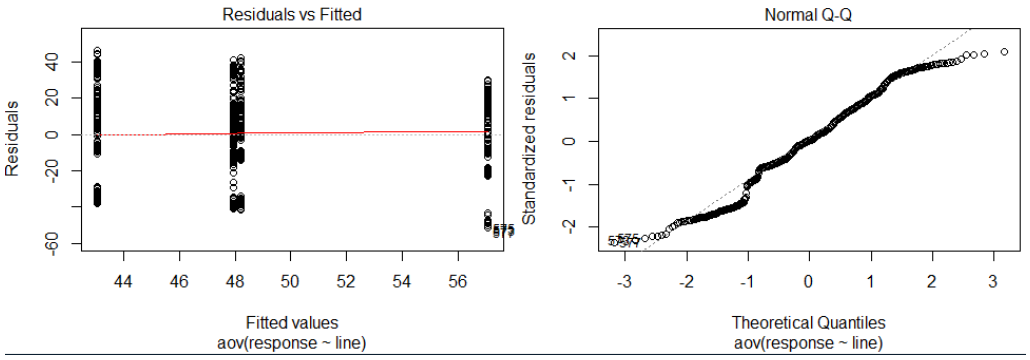
```
> summary(lineANOVA)
             Df Sum Sq Mean Sq F value   Pr(>F)
line          3  16038    5346   11.07 4.26e-07 ***
Residuals   664 320719     483
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

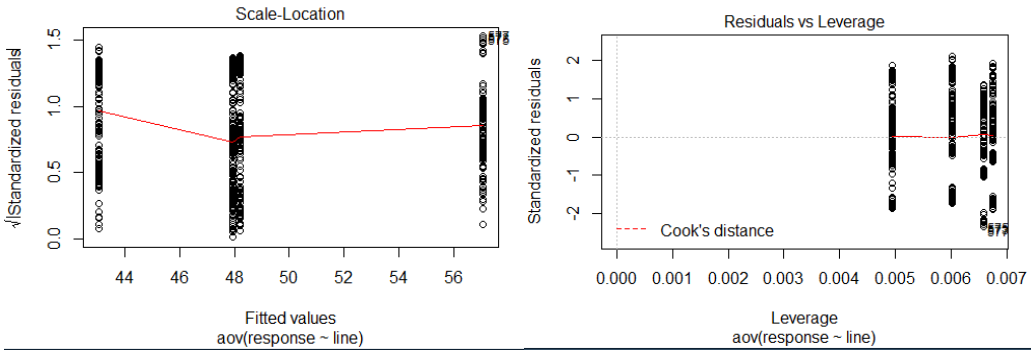
```
> summary(deviceANOVA)
             Df Sum Sq Mean Sq F value Pr(>F)
device        4 299921   74980    1350 <2e-16 ***
Residuals   663  36836      56
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Assumptions:

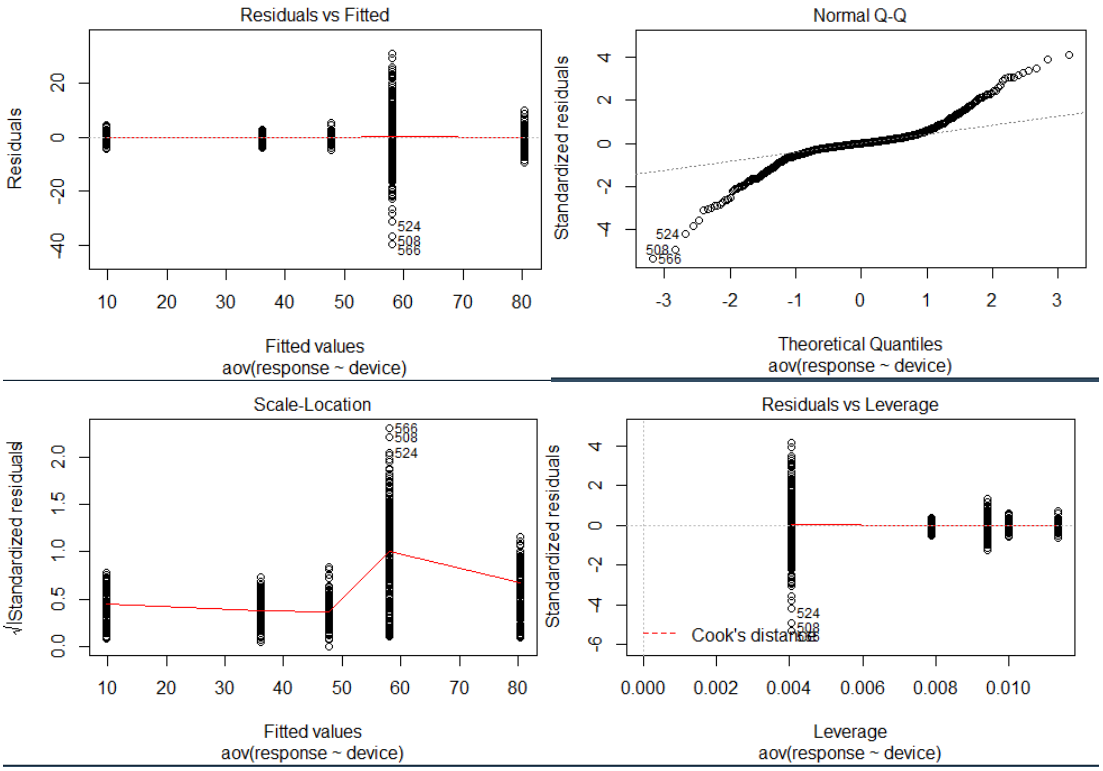
- All cases are independent.
- The residuals of the models are normal and small compared to the values
- The variances of each response group are similar (homoscedasticity)
- Data has no significant outliers

Residual plots of linear model for response time vs. assembly line:





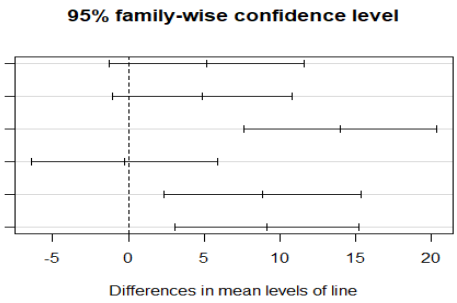
Residual plots of linear model for response time vs. device:



Pairwise Tests

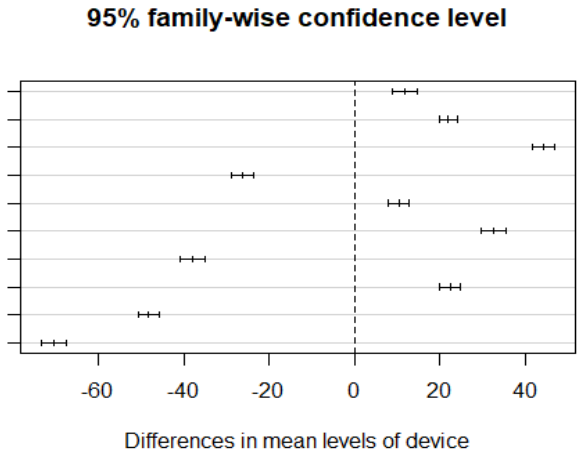
Tukey HSD multiple comparisons pairwise test for response time vs. assembly line:

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

Fit: aov(formula = response ~ line)

$line
                  diff       lwr       upr     p adj
green-blue    5.1579299 -1.241305 11.557164 0.1620493
orange-blue   4.8814287 -1.048402 10.811259 0.1477122
red-blue     14.0007354  7.646170 20.355300 0.0000001
orange-green -0.2765012 -6.401071  5.848069 0.9994376
red-green     8.8428055  2.306143 15.379468 0.0029501
red-orange    9.1193067  3.041425 15.197189 0.0007049
```



Tukey HSD multiple comparisons pairwise test for response time vs. device:

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

Fit: aov(formula = response ~ device)

$device
                                        diff        lwr       upr p adj
motor_controller2-motor_controller1  11.65822   8.830336  14.48611     0
pressure_switch-motor_controller1    21.94942  19.723180  24.17566     0
sequence_reader-motor_controller1    44.27699  41.594680  46.95930     0
timing_circuit-motor_controller1    -26.35350 -29.079322 -23.62768     0
pressure_switch-motor_controller2    10.29119   7.760039  12.82235     0
sequence_reader-motor_controller2    32.61877  29.678461  35.55908     0
timing_circuit-motor_controller2    -38.01173 -40.991777 -35.03167     0
sequence_reader-pressure_switch      22.32757  19.960172  24.69498     0
timing_circuit-pressure_switch      -48.30292 -50.719507 -45.88633     0
timing_circuit-sequence_reader      -70.63049 -73.472774 -67.78821     0
```



Discussion 1

Both groups of residual plots show that the data are (within reason) homoscedastic, with no outliers. They both look reasonably normal, but the normality of the residuals of the device to response time model is questionable. However, based on the extreme p-values of that model, the slight abnormality is probably acceptable.

The ANOVA models for both the device-to-response-time and line-to-response-time models show that there is a significant effect of both line ($p = 4.26 \cdot 10^{-7}$) and device type ($p < 2 \cdot 10^{-16}$) on response time. The effect of device type is much more significant, which can be seen easily in the boxplots.

Pairwise tests were very enlightening. They showed that all the assembly lines were about the same, except red was significantly slightly higher, and that all of the devices have significantly different response times.

However, we can't expect devices to have the same response times – that difference is natural. The line data was particularly skewed, since each line did not have the same number of data points for each device. For example, red had a disproportionate number of sequence readers, which had the highest response time and thus made red look slower. Thus, we must check for differences in lines between each device. I will assume the residual assumptions hold from the first tests – the data looks ok.

ANOVA:

```
> summary(ANOVA_motorController1)
                          Df Sum Sq Mean Sq F value Pr(>F)
line[log_motorController1]   3  13.64   4.546   2.153  0.097 .
Residuals                  123 259.69   2.111
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(ANOVA_motorController2)
                          Df Sum Sq Mean Sq F value Pr(>F)
line[log_motorController2]  3   1.69  0.5617   0.211  0.889
Residuals                  84 223.95  2.6661
```

```
> summary(ANOVA_pressureSwitch)
                        Df Sum Sq Mean Sq F value   Pr(>F)
line[log_pressureSwitch]   3   5687  1895.6   16.14 1.34e-09 ***
Residuals                243  28545   117.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

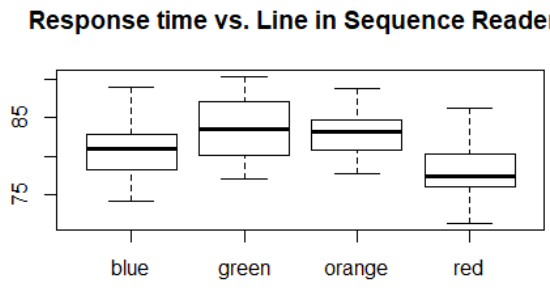
```
> summary(ANOVA_sequenceReader)
                        Df Sum Sq Mean Sq F value   Pr(>F)
line[log_sequenceReader]   3  496.2  165.40   14.75 4.78e-08 ***
Residuals                102 1144.1   11.22
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

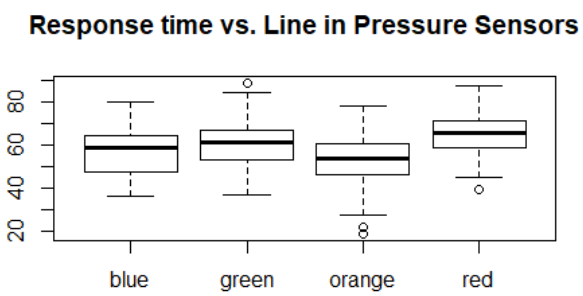
```
> summary(ANOVA_timingCircuit)
                        Df Sum Sq Mean Sq F value Pr(>F)
line[log_timingCircuit]  3    7.2   2.412   0.506  0.679
Residuals                96  457.4   4.765
```

With a confidence level of 95%, the only two devices that have significant differences between lines are pressure switches ($p = 1.34 \cdot 10^{-9}$) and sequence readers ($p = 4.78 \cdot 10^{-8}$). Now, we must run pairwise tests on these devices.

Pairwise test results:

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

Fit: aov(formula = aov(response[log_pressureSwitch] ~ line[log_pressureSwitch]))

$`line[log_pressureSwitch]`
                  diff        lwr        upr     p adj
green-blue    4.204279  -1.286502  9.6950607 0.1980863
orange-blue  -4.344076  -9.463059  0.7749074 0.1274254
red-blue      7.981309   2.220083 13.7425349 0.0022999
orange-green -8.548356 -13.175883 -3.9208277 0.0000181
red-green     3.777030  -1.552299  9.1063584 0.2602060
red-orange   12.325385   7.379977 17.2707930 0.0000000

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

Fit: aov(formula = aov(response[log_sequenceReader] ~ line[log_sequenceReader]))

$`line[log_sequenceReader]`
                  diff        lwr        upr     p adj
green-blue    2.4815144 -0.7126589  5.675688 0.1841392
orange-blue   1.8832287 -0.5124012  4.278859 0.1756720
red-blue     -3.0051442 -5.0962215 -0.914067 0.0016290
orange-green -0.5982856 -3.8907641  2.694193 0.9645084
red-green    -5.4866586 -8.5646409 -2.408676 0.0000571
red-orange   -4.8883729 -7.1267374 -2.650008 0.0000007
```





Device-line frequency table:

```
> table(device, line)
                 line
device            blue green orange red
  motor_controller1   25    32     40  30
  motor_controller2   16    24     28  20
  pressure_switch     45    62     90  50
  sequence_reader     30    10     24  42
  timing_circuit      50    20     20  10
```

Discussion 2

These tests are more enlightening. The pairwise tests show that the red line's pressures sensors are the slowest, and significantly slower than orange's or blue's lines. Orange's pressure sensors are the fastest, and are significantly faster than green's or red's. Surprisingly, red has by far the fastest sequence readers – significantly faster than any other line. Thus, the difference between lines must be mostly due to the distribution of devices. There is no way to confirm this, since green and orange have a very different device distribution but just happen to have similar average response times. More uniform data is necessary.

Conclusion

ANOVA models and pairwise tests provided good insight into the reasons for mistiming on the assembly lines. In conclusion, the two lines that are malfunctioning are probably blue and red – blue too fast and red too slow. The blue line may be faster in part because of its faster sequence readers (slowest device). However, the red line is only slower because it appears to have *more* sequence readers. To improve this analysis, uniform data should be collected with the same number of devices per line, or it should be clarified that the distribution actually represents the number of devices on each line and the lines are not the same.

```
########################################
# Dallin Romney - u1087199             #
# Design of Experiments Homework 4     #
# February 26, 2019                    #
########################################

rm(list = ls()) # Clear workspace
cat("\014")     # Clear console (control + L)

############## OPEN STATISTICS PROBLEMS ###############

#6.41
#predicted = as.factor(c(rep("book", 76), rep("print", 31), rep("online", 19)))
#actual    = as.factor(c(rep("book", 71), rep("print", 30), rep("online", 25)))
#chisq.test(df$Predicted, df$Actual, correct = "False")

############## ASSEMBLY LINE FAULT INVESTIGATION ##############

# Read in and organize data
assemblyLine = read.csv("C:/Users/Dallin/Google Drive/School/Design of Experiments/Homework/HW4/lineFaultData.csv")
response     = assemblyLine$response
line         = assemblyLine$line
device       = assemblyLine$device

# Visualize data
boxplot(response~line,   main = "Response time vs. Line")
boxplot(response~device, main = "Response time vs. Device")

# No outliers appear to be extreme, so I will include them

# Perform and view ANOVA analyses
lineANOVA   = aov(response~line)
deviceANOVA = aov(response~device)

summary(lineANOVA)
summary(deviceANOVA)

# show 4 residual plots for each, to verify assumptions
plot(lineANOVA)
plot(deviceANOVA)

# Perform and visualize pairwise test (Tukey's HSD test, multiple comparisons)
lineTukey   = TukeyHSD(lineANOVA,   conf.level = 0.95)
deviceTukey = TukeyHSD(deviceANOVA, conf.level = 0.95)

lineTukey
deviceTukey

plot(lineTukey)
plot(deviceTukey)

# Perform ANOVA on each device (line within device)
log_motorController1 = device == "motor_controller1"
log_motorController2 = device == "motor_controller2"
log_pressureSwitch   = device == "pressure_switch"
log_sequenceReader   = device == "sequence_reader"
log_timingCircuit    = device == "timing_circuit"

ANOVA_motorController1 = aov(aov(response[log_motorController1]~line[log_motorController1]))
ANOVA_motorController2 = aov(aov(response[log_motorController2]~line[log_motorController2]))
ANOVA_pressureSwitch   = aov(aov(response[log_pressureSwitch  ]~line[log_pressureSwitch  ]))
ANOVA_sequenceReader   = aov(aov(response[log_sequenceReader  ]~line[log_sequenceReader  ]))
ANOVA_timingCircuit    = aov(aov(response[log_timingCircuit   ]~line[log_timingCircuit   ]))

summary(ANOVA_motorController1)
summary(ANOVA_motorController2)
summary(ANOVA_pressureSwitch)
summary(ANOVA_sequenceReader)
summary(ANOVA_timingCircuit)

# Visualize devices with significant differences between lines
boxplot(response[log_pressureSwitch]~line[log_pressureSwitch], main = "Response time vs. Line in Pressure Sensors")
boxplot(response[log_sequenceReader]~line[log_sequenceReader], main = "Response time vs. Line in Sequence Readers")

# Perform pairwise tests on devices with significant differences between lines
pressureSensorTukey = TukeyHSD(ANOVA_pressureSwitch, conf.level = 0.95)
sequenceReaderTukey = TukeyHSD(ANOVA_sequenceReader, conf.level = 0.95)

pressureSensorTukey
sequenceReaderTukey

# Count device frequencies
table(device, line)
```