MAT325 Project 5

Jordan Badstuebner

19 August 2020

```
#############
####  1  ####
#############
```

(a) Estimate the missing value.

```
##   ï..Farm Surface Trickle CenterPivot Lateral Subirrigation
## 1       1      NA     248         391     423           350
## 2       2     636     382         434     461           370
## 3       3     591     348         492     504           460
## 4       4     603     366         468     580           452
## 5       5     649     258         457     449           343
## 6       6     512     321         406     464           340
```

We can see observe immediately from the head of the data that there exists a missing value in the Surface vector. Let's verify all casess of NA values in the dataset:

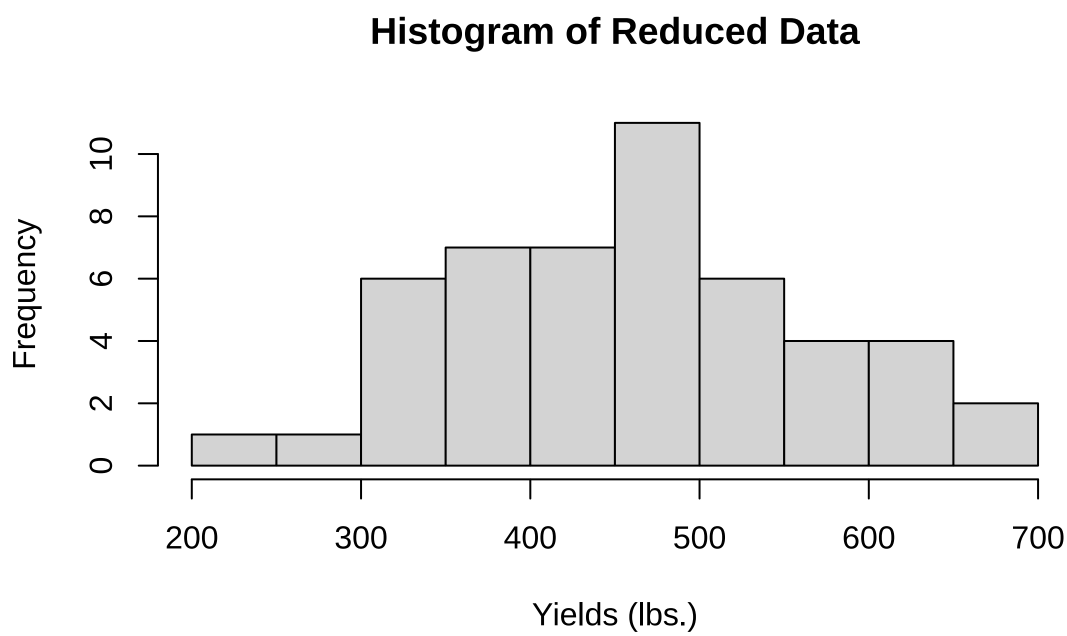
First, let's rearrange the data such that it will be readable for our linear model code.

```
#CHECK FOR ALL MISSING DATA IN THE DATASET
df1_m[!complete.cases(df1_m),]
```

```
##   ï..Farm variable value
## 1       1  Surface    NA
```

From the complete cases funtion we observe that row 1 of the "value" vector is the only case of missing data in the data set. We can use analysis of variance to estimate the missing value.

Fist, let's make sure the data meets the assumptions for our analysis of variance.



From the histogram, the data appears to be nomally distributed. We will also assume the observations are independent through good experimental designs. Now let's run the ANOVA.

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## df1_m$ï..Farm    1  32829   32829   13.26 0.000725 ***
## df1_m$variable   4 392087   98022   39.58  6.9e-14 ***
## Residuals       43 106490    2477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

Now let's run the predictions.

```
#PREDICT NEW DATA
predict(res.aov, newdata = df1_m)
```

```
##        1        2        3        4        5        6        7        8
## 581.5150 588.9009 596.2868 603.6726 611.0585 618.4444 625.8303 633.2162
##        9       10       11       12       13       14       15       16
## 640.6021 647.9880 319.9635 327.3494 334.7353 342.1212 349.5071 356.8929
##       17       18       19       20       21       22       23       24
## 364.2788 371.6647 379.0506 386.4365 431.1635 438.5494 445.9353 453.3212
##       25       26       27       28       29       30       31       32
## 460.7071 468.0929 475.4788 482.8647 490.2506 497.6365 464.3635 471.7494
##       33       34       35       36       37       38       39       40
## 479.1353 486.5212 493.9071 501.2929 508.6788 516.0647 523.4506 530.8365
##       41       42       43       44       45       46       47       48
## 356.4635 363.8494 371.2353 378.6212 386.0071 393.3929 400.7788 408.1647
##       49       50
## 415.5506 422.9365
```

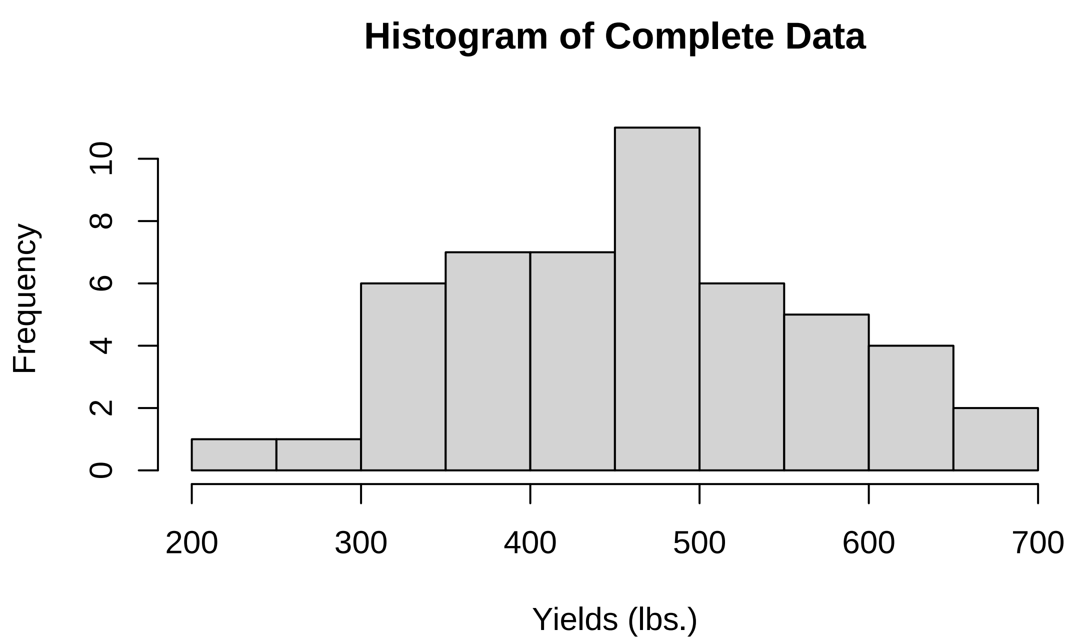
Our missing value is estimated to be 581.5150.

(b) Analyze the data by replacing the missing value with the estimate obtained in part (a) and then perform an analysis of variance.

```
#REPLACE THE MSSING VALUE WITH THE ESTIMATED VALUE

#INPUT PREDICTED VALUE INTO DATAFRAME
df1_m[1,3] = 581.5150
head(df1_m)
```

```
##   ï..Farm variable   value
## 1       1  Surface 581.515
## 2       2  Surface 636.000
## 3       3  Surface 591.000
## 4       4  Surface 603.000
## 5       5  Surface 649.000
## 6       6  Surface 512.000
```



From the histogram, the new data appears to be nomally distributed. We will also assume the observations are independent through good experimental designs. Now let's run the ANOVA.

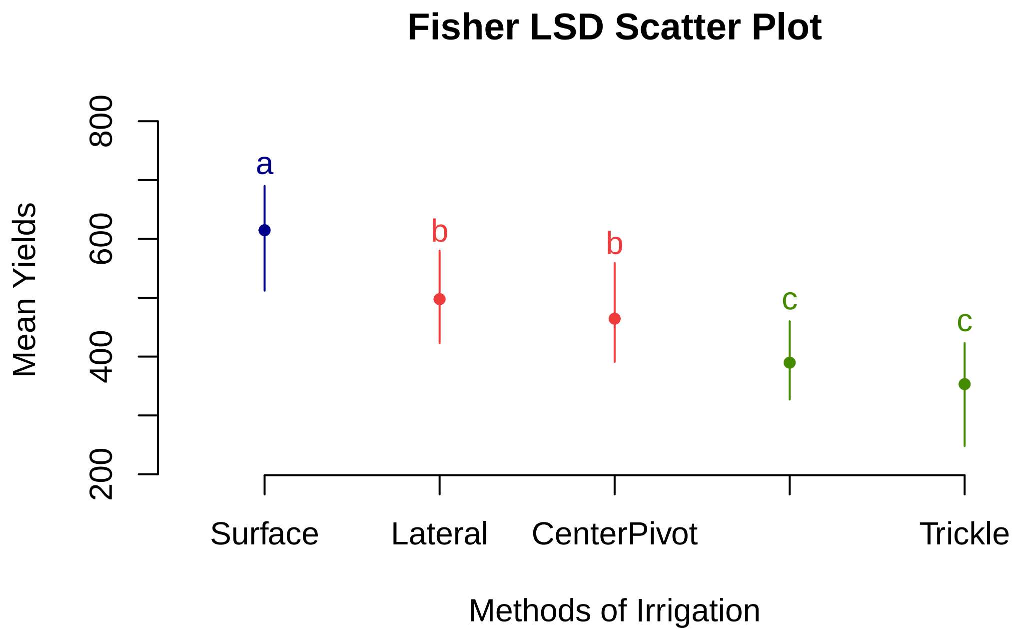
```
## Analysis of Variance Table
##
## Response: df1_m$value
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## df1_m$ï..Farm    1  22502   22502  9.2977  0.003876 **
## df1_m$variable  4 416522  104131 43.0253 1.151e-14 ***
## Residuals      44 106490    2420
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c) Is there a significant difference in the mean yields for the different methods of irrigation? Use $\alpha = 0.05$.

With $p - val = 1.151 \times 10^{-14} < \alpha = 0.05$, there is sufficient evidence to suggest at a significance level of 0.05 that there is a significant difference in the mean yields for the different methods of irrigation.

(d) Use the least significant difference criterion to identify which pairs of methods of irrigation have significantly different mean yields.

```
## $statistics
##   MSerror Df     Mean       CV  t.value      LSD
##      2420 44 463.9303 10.60364 2.015368 44.33809
##
## $parameters
##         test p.ajusted         name.t ntr alpha
##   Fisher-LSD      none df1_m$variable   5  0.05
##
## $means
##               df1_m$value      std  r      LCL      UCL Min Max    Q25   Q50
## CenterPivot      464.4000 48.23369 10 433.0482 495.7518 391 559 439.75 467.0
## Lateral          497.6000 52.00684 10 466.2482 528.9518 423 580 461.75 486.5
## Subirrigation    389.7000 52.73635 10 358.3482 421.0518 327 460 344.75 374.0
## Surface          614.7515 53.67146 10 583.3997 646.1033 512 690 588.75 605.5
## Trickle          353.2000 60.32836 10 321.8482 384.5518 248 423 327.75 373.0
##                  Q75
## CenterPivot   486.25
## Lateral       544.00
## Subirrigation 443.75
## Surface       645.75
## Trickle       395.50
##
## $comparison
## NULL
##
## $groups
##               df1_m$value groups
## Surface          614.7515      a
## Lateral          497.6000      b
## CenterPivot      464.4000      b
## Subirrigation    389.7000      c
## Trickle          353.2000      c
##
## attr(,"class")
## [1] "group"
```



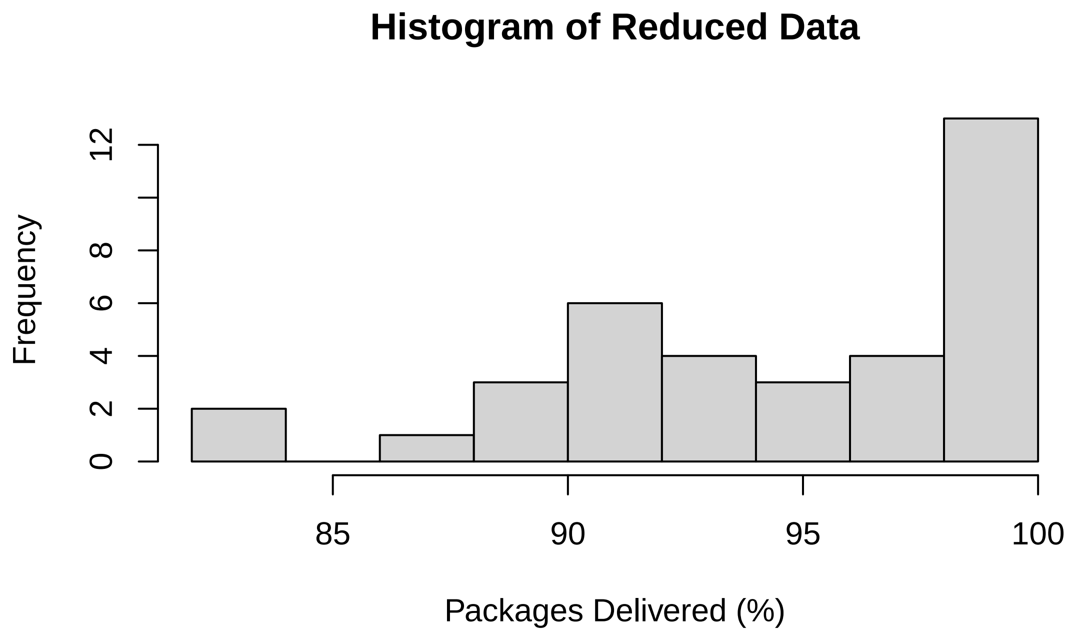
The Fisher LSD Scatter Plot above represents which pairs of methods of irrigation have significantly different mean yields. The missing method is "Subirrigation".

(e) Obtain the sum of squares for an ANOVA table by fitting complete and reduced models using a statistical software program.

From the 2 ANOVA tables above,

1. Reduced Model
   (a) FARM SST = 32829
   (b) METHOD SST = 392087
   (c) SSE = 106490
2. Complete Model
   (a) FARM SST = 22502
   (b) METHOD SST = 416522
   (c) SSE = 106490

(a) Obtain the sum of squares for an ANOVA table by fitting complete and reduced models using a statistical software program.



We see here that the data is no where near normally distrubuted. This contradicts the assumptions of ANOVA. We will proceed with the analysis, but we may be increasing the chance of a false positive result.

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)    
## df2_m$ï..Method   3  270.2   90.06  15.841 8.38e-06 ***
## df2_m$variable    9  386.4   42.94   7.553 4.42e-05 ***
## Residuals        23  130.8    5.68                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 4 observations deleted due to missingness
```

Below are predictions for the new data based on the ANOVA:

```
#PREDICT NEW DATA
predict(res.aov_2, newdata = df2_m)
```

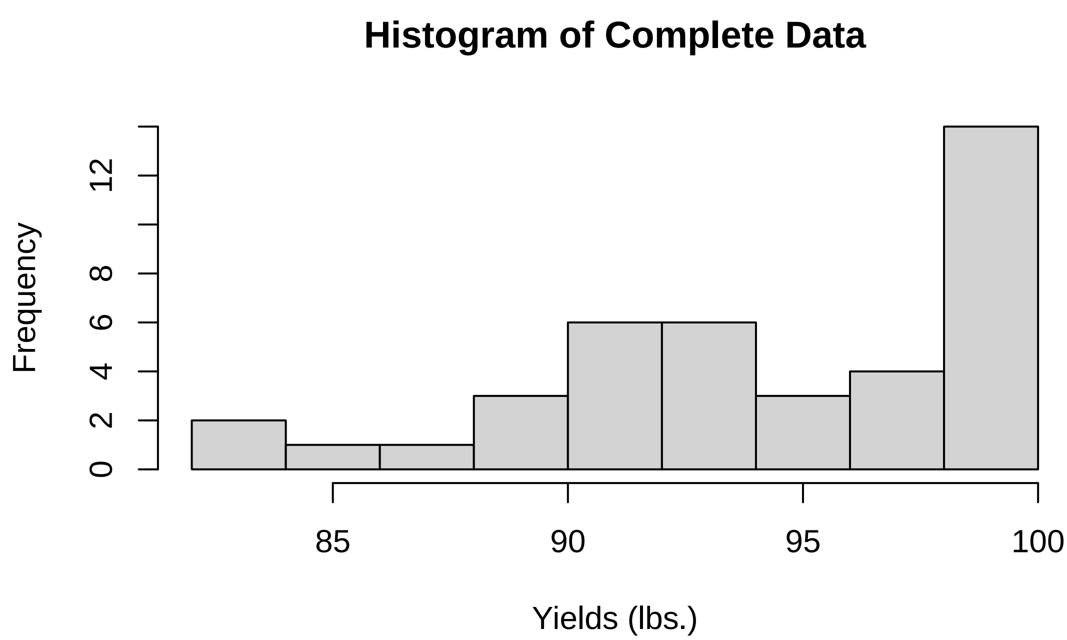
```
##         1         2         3         4         5         6         7         8 
##  85.37635  89.21108  92.07908  92.90985  92.80726  96.64199  99.50999 100.34076 
##         9        10        11        12        13        14        15        16 
##  85.77726  89.61199  92.47999  93.31076  90.20726  94.04199  96.90999  97.74076 
##        17        18        19        20        21        22        23        24 
##  94.45726  98.29199 101.15999 101.99076  90.40726  94.24199  97.10999  97.94076 
##        25        26        27        28        29        30        31        32 
##  94.15418  97.98891 100.85691 101.68768  85.88226  89.71699  92.58499  93.41576 
##        33        34        35        36        37        38        39        40 
##  87.50726  91.34199  94.20999  95.04076  92.37635  96.21108  99.07908  99.90985
```

```
#REPLACE THE MSSING VALUE WITH THE ESTIMATED VALUE
```

```r
#INPUT PREDICTED VALUE INTO DATAFRAME
df2_m[1,3] = 85.37635
df2_m[11,3] = 92.47999
df2_m[28,3] = 100
df2_m[37,3] = 92.37635
head(df2_m)
```

```
##   ï..Method variable    value
## 1        M1       C1 85.37635
## 2        M2       C1 87.10000
## 3        M3       C1 91.60000
## 4        M4       C1 95.50000
## 5        M1       C2 90.20000
## 6        M2       C2 99.50000
```

NOTE: Many of the predictions returned as greater than 100%. Since 100% is the upper bound, all values returning as greater than 100% were input into the data set at 100% to satisfy the bound of percentage.



Observe that the corrected data is is no where near normally distrubuted. This contradicts the assumptions of ANOVA. We will proceed with the analysis, but we may be increasing the chance of a false positive result.

```
## Analysis of Variance Table
##
## Response: df2_m$value
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## df2_m$ï..Method  3 337.49 112.496 22.8931 1.398e-07 ***
## df2_m$variable   9 438.79  48.754  9.9215 1.552e-06 ***
## Residuals       27 132.68   4.914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the 2 ANOVA tables above,

1. Reduced Model
   (a) METHOD SST = 270.2
   (b) CITY SST = 386.4
   (c) SSE = 130.8
2. Complete Model
   (a) METHOD SST = 337.49
   (b) CITY SST = 438.79
   (c) SSE = 132.68

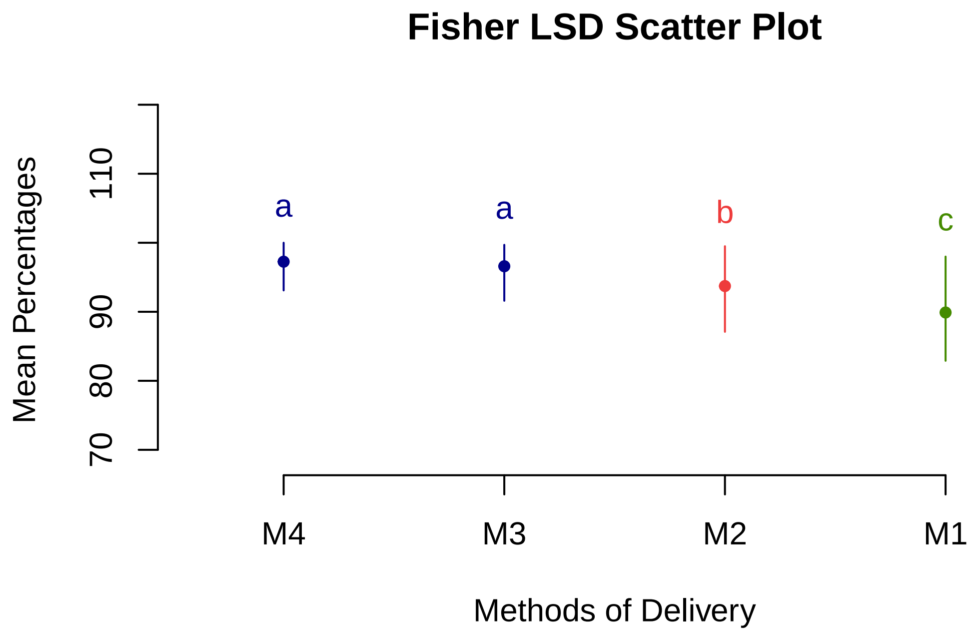
(b) Is there significant evidence of a difference in the four methods of delivery based on the percentage of packages delivered within five days?

With $p - val = 1.398 \times 10^{-7} < \alpha = 0.05$, there is sufficient evidence to suggest, at a significance level of 0.05, that there is a significant difference in the four methods of delivery based on the percentage of packages delivered within five days.

NOTE: Again, the LM and ANOVA were both run with non-normally distributed data.

(c) Use the least significant difference criterion to identify which pairs of methods of delivery have significantly different mean percentages.

```
## $statistics
##   MSerror Df     Mean       CV  t.value      LSD
##     4.914 27 94.37082 2.348983 2.051831 2.034108
##
## $parameters
##         test p.ajusted          name.t ntr alpha
##   Fisher-LSD      none df2_m$ï..Method   4  0.05
##
## $means
##    df2_m$value      std  r      LCL      UCL  Min   Max      Q25   Q50      Q75
## M1    89.89527 5.178326 10 88.45694 91.33360 82.9  98.0 86.18226 89.80 92.15726
## M2    93.73000 4.433221 10 92.29167 95.16833 87.1  99.5 91.42500 92.35 97.60000
## M3    96.59800 3.040184 10 95.15967 98.03633 91.6  99.7 94.12500 98.15 98.97500
## M4    97.26000 2.790141 10 95.82167 98.69833 93.1 100.0 94.45000 98.75 99.37500
##
## $comparison
## NULL
##
## $groups
##    df2_m$value groups
## M4    97.26000      a
## M3    96.59800      a
## M2    93.73000      b
## M1    89.89527      c
##
## attr(,"class")
## [1] "group"
```



The Fisher LSD Scatter Plot above represents which pairs of methods of delivery have significantly different mean percentages.