

Homework#7

uden

10/10/2018

Go back to your “thinking on paper” exercise, and decide on a pattern that you might expect in your experiment if a specific hypothesis were true.

Since the “thinking on paper” exercise, my direction with Jane has changed slightly. Now, we are collecting data on flowering *Centaurea* in a common garden. Date of first flower will be recorded from individuals taken from Norway, Germany, Luxembourg, Switzerland, France and Spain. I don’t know much about the species at all, but for today’s purpose, I’m going to hypothesise that flowering time is extended the further south you move along the species’ range -growing season decreases as you move north. So, the following order of flowering dates would be predicted for each population:

Norway: earliest flower Germany Luxembourg Switzerland France Spain: latest flower

Since this is a common garden experiment, variation between flowering times (phenotypic variation) is due to genetic variation, not environmental. If there is no significant variation, then flowering time is dependent on environmental variation.

To start, simply assume that the data in each of your treatment groups follow a normal distribution. Specify the sample sizes, means, and variances for each group that would be reasonable if your hypothesis were true.

Instead of date of first flower, I’m going to say date since first individual flowered. So, an individual with time 0 was the first one to flower. An individual with time 2 flowered 2 days later. Jane reckons this will take about six weeks of recording. So, I’ll split the six weeks into six -most of Norway flowers around the first 7 days, most of Germany flowers after 14 days, Luxembourg’s flowers after 21 days and so on. We will start with a sample size of 10 for each treatment. Variance will be 5 for all of them.

Using the methods we have covered in class, write code to create a random data set that has these attributes. Organize these data into a data frame or tibble with the appropriate structure.

```
library(ggplot2)
library(MASS)

#Data frame construction for ANOVA

nGroup <- 6
nName <- c("Norway", "Germany", "Luxembourg", "Switzerland", "France", "Spain")
nSize <- c(10, 10, 10, 10, 10, 10) #how many sample we will have in each group. "rule of ten:" at least
nMean <- c(7, 14, 21, 28, 35, 42) #mean of each group. this is point at which we can simulate what data
nSD <- c(5, 5, 5, 5, 5, 5) #standard deviation. variation between a data point and the means of treatment

ID <- 1:(sum(nSize)) #generate row in data set for each observation

#start exploration with NORMAL DISTRIBUTION:

Days <- c(rnorm(n=nSize[1], mean=nMean[1], sd=nSD[1])), #this will generate 10 random variates, with a mean
```

```

    rnorm(n=nSize[2], mean=nMean[2], sd=nSD[2]),
    rnorm(n=nSize[3], mean=nMean[3], sd=nSD[3]),
    rnorm(n=nSize[4], mean=nMean[4], sd=nSD[4]),
    rnorm(n=nSize[5], mean=nMean[5], sd=nSD[5]),
    rnorm(n=nSize[6], mean=nMean[6], sd=nSD[6]))
Country <- rep(nName, nSize)
table(Country)

## Country
##      France      Germany Luxembourg      Norway      Spain Switzerland
##          10          10          10          10          10          10

#string to gether into data frame:
ANOfdata <- data.frame(ID, Country, Days)
str(ANOfdata)

## 'data.frame':    60 obs. of  3 variables:
## $ ID      : int  1 2 3 4 5 6 7 8 9 10 ...
## $ Country: Factor w/ 6 levels "France","Germany",...: 4 4 4 4 4 4 4 4 4 4 ...
## $ Days   : num  16.92 11.45 7.66 7.32 1.79 ...

head(ANOfdata)

##   ID Country      Days
## 1  1  Norway 16.917224
## 2  2  Norway 11.453971
## 3  3  Norway  7.656224
## 4  4  Norway  7.321276
## 5  5  Norway  1.788539
## 6  6  Norway  1.513608

```

Now write code to analyze the data (probably as an ANOVA or regression analysis, but possibly as a logistic regression or contingency table analysis. Write code to generate a useful graph of the data.

```

# Basic ANOVA model in R
ANOfmodel <- aov(Days~Country, data=ANOfdata) #"analysis of variance" response is function of (~) predict
print(ANOfmodel)

## Call:
##   aov(formula = Days ~ Country, data = ANOfdata)
##
## Terms:
##              Country Residuals
## Sum of Squares 8620.265 1555.096
## Deg. of Freedom      5      54
##
## Residual standard error: 5.366384
## Estimated effects may be unbalanced

print(summary(ANOfmodel))

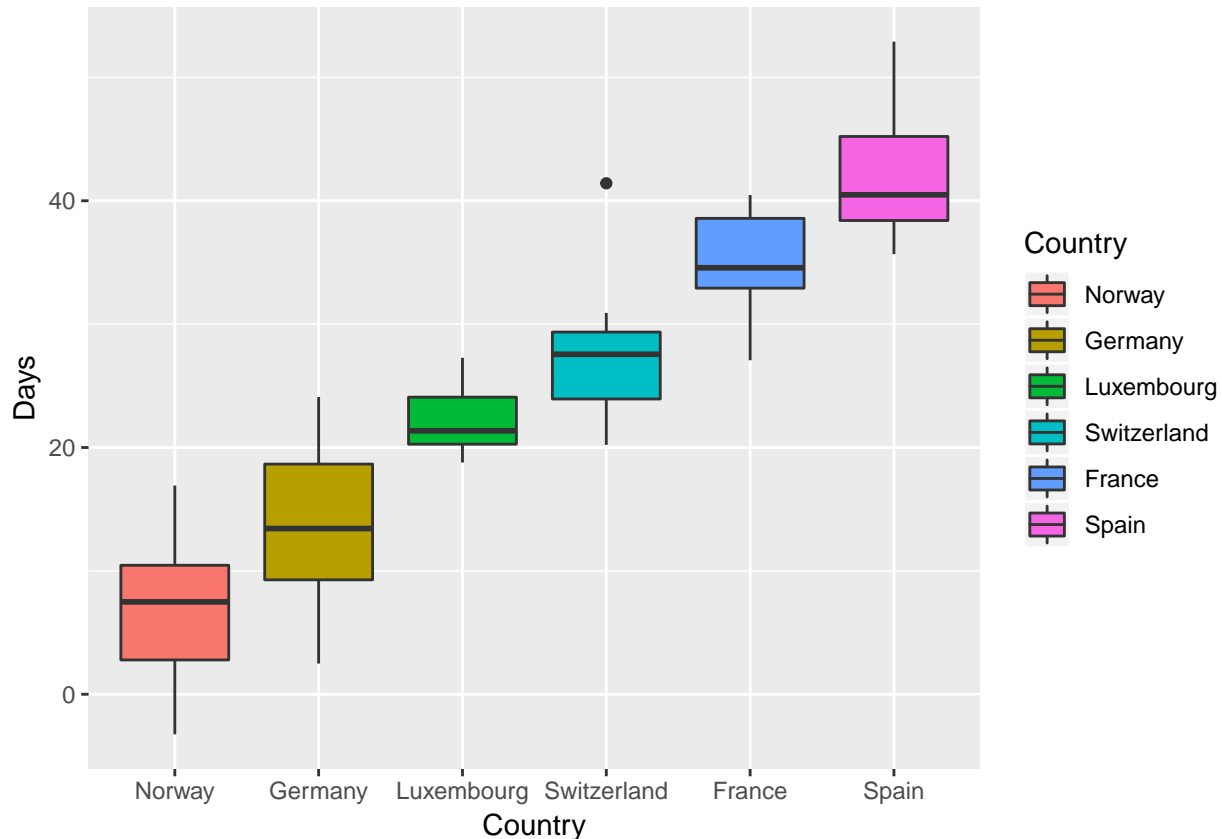
##              Df Sum Sq Mean Sq F value Pr(>F)
## Country        5   8620   1724.1    59.87 <2e-16 ***
## Residuals     54   1555     28.8

```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#I want my countries to be plotted in latitudinal order:
ANOdata$Country = with(ANOdata, reorder(Country, ID))

ANOplot <- ggplot(data=ANOdata,
                  aes(x=Country, y=Days,
                     fill= Country)) +
  geom_boxplot()
print(ANOplot)
```



Try running your analysis multiple times to get a feeling for how variable the results are with the same parameters, but different sets of random numbers.

Now begin adjusting the means of the different groups. Given the sample sizes you have chosen, how small can the differences between the groups be (the “effect size”) for you to still detect a significant pattern ($p < 0.05$)?

Lets say there isn't a one week gap between first flowers of each country. Instead, lets try 5 days:

```
#Data frame construction for ANOVA
nGroup <- 6
nName <- c("Norway", "Germany", "Luxembourg", "Switzerland", "France", "Spain")
nSize <- c(10, 10, 10, 10, 10, 10)
nMean <- c(5, 10, 15, 20, 25, 30) #changed the means from 7 to 5 day intervals.
nSD <- c(5, 5, 5, 5, 5, 5)
```

```

ID <- 1:(sum(nSize))
Days <- c(rnorm(n=nSize[1], mean=nMean[1], sd=nSD[1]),
          rnorm(n=nSize[2], mean=nMean[2], sd=nSD[2]),
          rnorm(n=nSize[3], mean=nMean[3], sd=nSD[3]),
          rnorm(n=nSize[4], mean=nMean[4], sd=nSD[4]),
          rnorm(n=nSize[5], mean=nMean[5], sd=nSD[5]),
          rnorm(n=nSize[6], mean=nMean[6], sd=nSD[6]))
Country <- rep(nName, nSize)
table(Country)

## Country
##      France      Germany Luxembourg      Norway      Spain Switzerland
##           10           10           10           10           10           10

ANOfdata <- data.frame(ID, Country, Days)
str(ANOfdata)

## 'data.frame':    60 obs. of  3 variables:
##  $ ID      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Country: Factor w/ 6 levels "France","Germany",...: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Days    : num  9.33 4.68 3.74 6.37 1.19 ...

head(ANOfdata)

##   ID Country    Days
## 1  1  Norway 9.332638
## 2  2  Norway 4.678366
## 3  3  Norway 3.735194
## 4  4  Norway 6.374121
## 5  5  Norway 1.185130
## 6  6  Norway 2.990520

# Basic ANOVA model in R
ANOfmodel <- aov(Days~Country, data=ANOfdata)
print(ANOfmodel)

## Call:
## aov(formula = Days ~ Country, data = ANOfdata)
##
## Terms:
##              Country Residuals
## Sum of Squares 4658.867 1206.921
## Deg. of Freedom      5      54
##
## Residual standard error: 4.72762
## Estimated effects may be unbalanced

print(summary(ANOfmodel))

##              Df Sum Sq Mean Sq F value Pr(>F)
## Country        5   4659    931.8   41.69 <2e-16 ***
## Residuals      54   1207     22.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

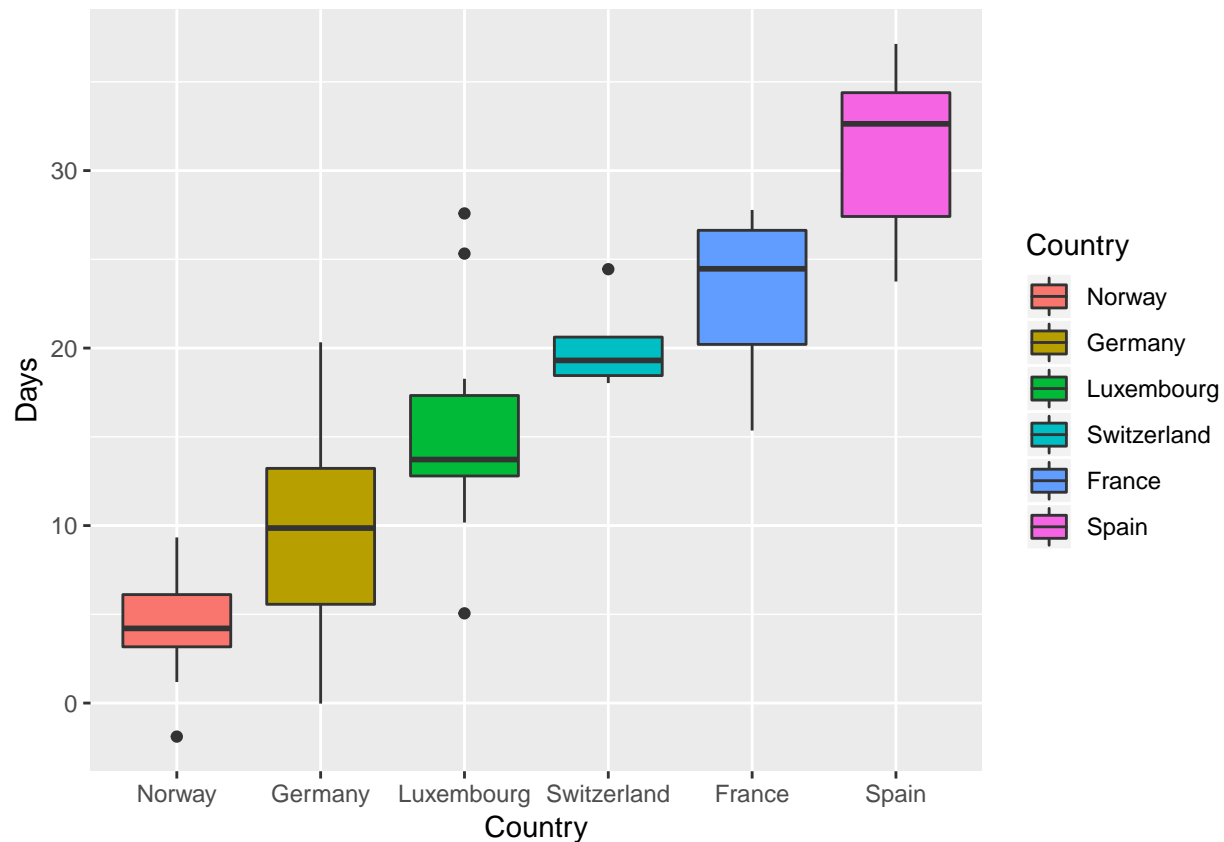
ANOfdata$Country = with(ANOfdata, reorder(Country, ID))
ANOfplot <- ggplot(data=ANOfdata,

```

```

aes(x=Country, y=Days,
    fill= Country)) +
geom_boxplot()
print(ANOPlot)

```



-this still shows a significant pattern. Lets try a mean difference of 2 days.

```

#Data frame construction for ANOVA
nGroup <- 6
nName <- c("Norway", "Germany", "Luxembourg", "Switzerland", "France", "Spain")
nSize <- c(10, 10, 10, 10, 10, 10)
nMean <- c(2, 4, 6, 8, 10, 12) # change mean flowering time to 2 day difference
nSD <- c(5, 5, 5, 5, 5, 5)
ID <- 1:(sum(nSize))
Days <- c(rnorm(n=nSize[1], mean=nMean[1], sd=nSD[1]),
          rnorm(n=nSize[2], mean=nMean[2], sd=nSD[2]),
          rnorm(n=nSize[3], mean=nMean[3], sd=nSD[3]),
          rnorm(n=nSize[4], mean=nMean[4], sd=nSD[4]),
          rnorm(n=nSize[5], mean=nMean[5], sd=nSD[5]),
          rnorm(n=nSize[6], mean=nMean[6], sd=nSD[6]))
Country <- rep(nName, nSize)
table(Country)

```

```

## Country
##      France      Germany  Luxembourg      Norway      Spain Switzerland
##          10           10           10           10           10           10

```

```

ANOfdata <- data.frame(ID, Country, Days)
str(ANOfdata)

## 'data.frame':    60 obs. of  3 variables:
## $ ID      : int  1 2 3 4 5 6 7 8 9 10 ...
## $ Country: Factor w/ 6 levels "France","Germany",...: 4 4 4 4 4 4 4 4 4 4 ...
## $ Days    : num  0.17 7.486 9.881 6.621 -0.178 ...

head(ANOfdata)

##   ID Country      Days
## 1  1  Norway  0.1696603
## 2  2  Norway  7.4861015
## 3  3  Norway  9.8812140
## 4  4  Norway  6.6212792
## 5  5  Norway -0.1783430
## 6  6  Norway  7.6211712

# Basic ANOVA model in R
ANOfmodel <- aov(Days~Country, data=ANOfdata)
print(ANOfmodel)

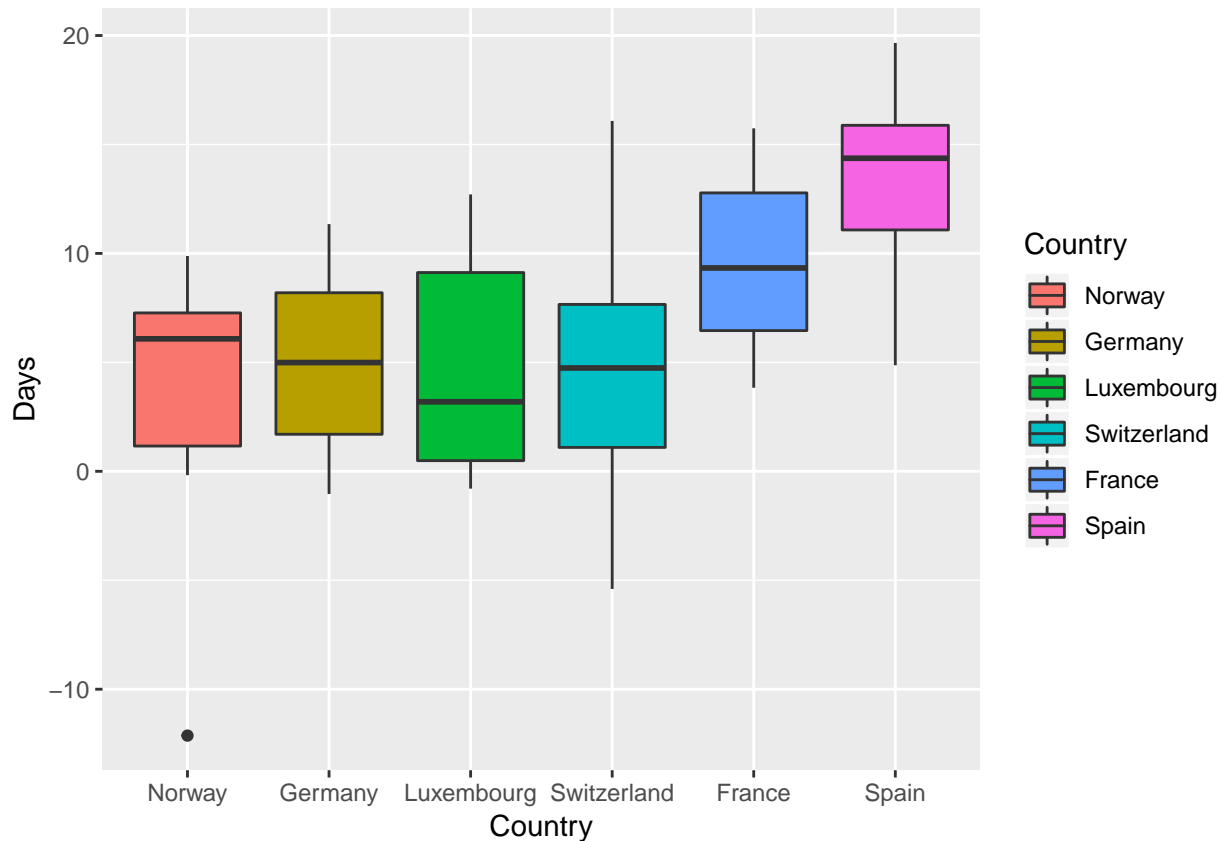
## Call:
## aov(formula = Days ~ Country, data = ANOfdata)
##
## Terms:
##                Country Residuals
## Sum of Squares    725.0698 1452.0726
## Deg. of Freedom         5         54
##
## Residual standard error: 5.185579
## Estimated effects may be unbalanced

print(summary(ANOfmodel))

##                Df Sum Sq Mean Sq F value    Pr(>F)
## Country          5   725.1   145.01    5.393 0.00043 ***
## Residuals       54 1452.1    26.89
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ANOfdata$Country = with(ANOfdata, reorder(Country, ID))
ANOfplot <- ggplot(data=ANOfdata,
                  aes(x=Country, y=Days,
                      fill= Country)) +
  geom_boxplot()
print(ANOfplot)

```



-After running the code 32 times, a non significant relationship between country of origina and flowering time was found. So, lets try adding one day to the mean.

#Data frame construction for ANOVA

```
nGroup <- 6
```

```
nName <- c("Norway", "Germany", "Luxembourg", "Switzerland", "France", "Spain")
```

```
nSize <- c(10, 10, 10, 10, 10, 10)
```

```
nMean <- c(3, 6, 9, 12, 15, 18) # change mean flowering time to 3 day difference
```

```
nSD <- c(5, 5, 5, 5, 5, 5)
```

```
ID <- 1:(sum(nSize))
```

```
Days <- c(rnorm(n=nSize[1], mean=nMean[1], sd=nSD[1]),
          rnorm(n=nSize[2], mean=nMean[2], sd=nSD[2]),
          rnorm(n=nSize[3], mean=nMean[3], sd=nSD[3]),
          rnorm(n=nSize[4], mean=nMean[4], sd=nSD[4]),
          rnorm(n=nSize[5], mean=nMean[5], sd=nSD[5]),
          rnorm(n=nSize[6], mean=nMean[6], sd=nSD[6]))
```

```
Country <- rep(nName, nSize)
```

```
table(Country)
```

```
## Country
```

```
##      France      Germany  Luxembourg      Norway      Spain Switzerland
```

```
##          10          10          10          10          10          10
```

```
ANOfdata <- data.frame(ID, Country, Days)
```

```
str(ANOfdata)
```

```
## 'data.frame': 60 obs. of 3 variables:
## $ ID : int 1 2 3 4 5 6 7 8 9 10 ...
## $ Country: Factor w/ 6 levels "France","Germany",...: 4 4 4 4 4 4 4 4 4 4 ...
## $ Days : num 1.99 3.11 4.92 -0.85 -4.89 ...
```

```
head(ANodata)
```

```
## ID Country Days
## 1 1 Norway 1.9878936
## 2 2 Norway 3.1106715
## 3 3 Norway 4.9169890
## 4 4 Norway -0.8501798
## 5 5 Norway -4.8948289
## 6 6 Norway 7.4537088
```

```
# Basic ANOVA model in R
```

```
ANOModel <- aov(Days~Country, data=ANodata)
print(ANOModel)
```

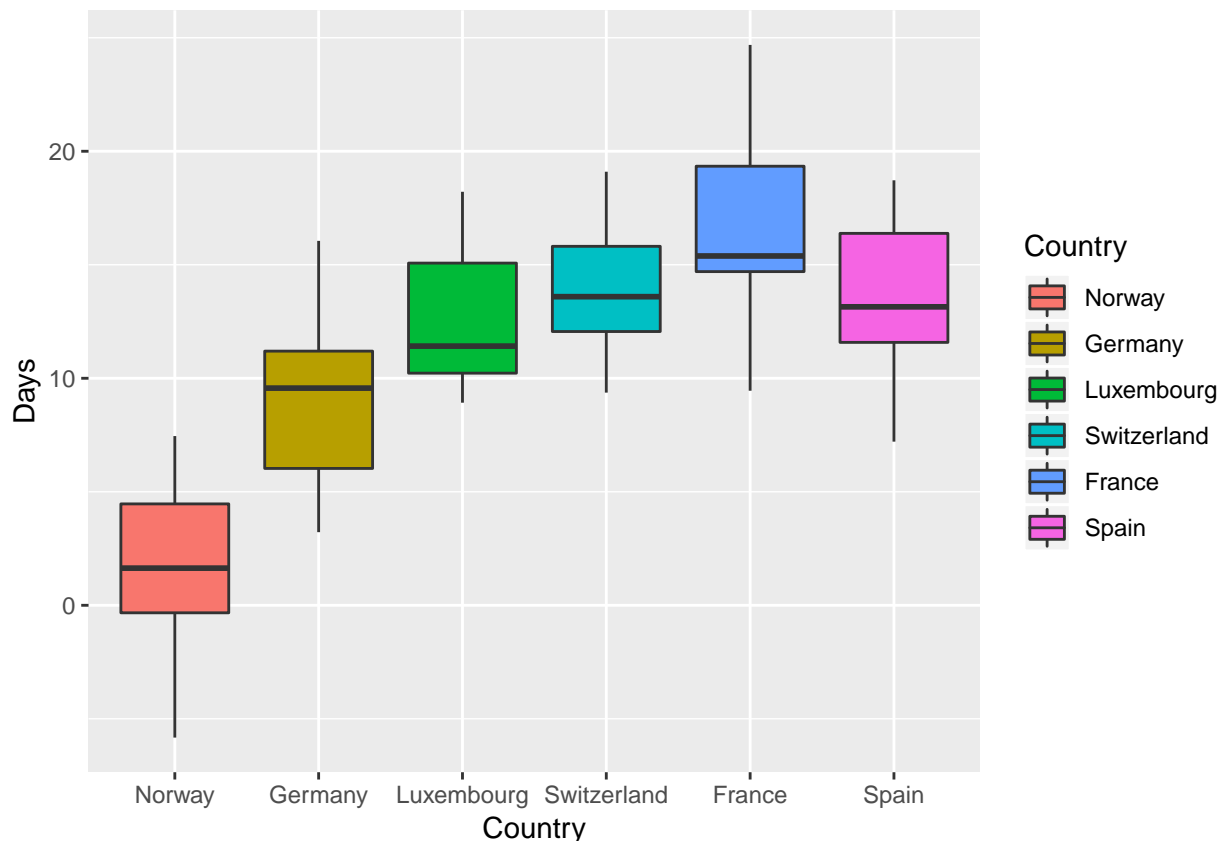
```
## Call:
## aov(formula = Days ~ Country, data = ANodata)
##
## Terms:
## Country Residuals
## Sum of Squares 1447.7105 804.9656
## Deg. of Freedom 5 54
##
## Residual standard error: 3.860929
## Estimated effects may be unbalanced
```

```
print(summary(ANOModel))
```

```
## Df Sum Sq Mean Sq F value Pr(>F)
## Country 5 1448 289.54 19.42 5.14e-11 ***
## Residuals 54 805 14.91
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ANodata$Country = with(ANodata, reorder(Country, ID))
```

```
ANOPlot <- ggplot(data=ANodata,
                  aes(x=Country, y=Days,
                      fill= Country)) +
  geom_boxplot()
print(ANOPlot)
```

There seems to be a consistent pattern of significance, even after running the code many times. So, I would not conclude that there is a significant relationship between country and flowering time unless the difference between means is at least 3 days.

Alternatively, for the effect sizes you originally hypothesized, what is the minimum sample size you would need in order to detect a statistically significant effect? Again, run the model a few times with the same parameter set to get a feeling for the effect of random variation in the data.

I'll start by halving the sample size:

```
#Data frame construction for ANOVA
nGroup <- 6
nName <- c("Norway", "Germany", "Luxembourg", "Switzerland", "France", "Spain")
nSize <- c(5, 5, 5, 5, 5, 5) #changed sample size from 10 to 5 plants per country.
nMean <- c(3, 6, 9, 12, 15, 18)
nSD <- c(5, 5, 5, 5, 5, 5)
ID <- 1:(sum(nSize))
Days <- c(rnorm(n=nSize[1], mean=nMean[1], sd=nSD[1]),
          rnorm(n=nSize[2], mean=nMean[2], sd=nSD[2]),
          rnorm(n=nSize[3], mean=nMean[3], sd=nSD[3]),
          rnorm(n=nSize[4], mean=nMean[4], sd=nSD[4]),
          rnorm(n=nSize[5], mean=nMean[5], sd=nSD[5]),
          rnorm(n=nSize[6], mean=nMean[6], sd=nSD[6]))
Country <- rep(nName, nSize)
table(Country)
```

```
## Country
```

```
##      France      Germany Luxembourg      Norway      Spain Switzerland
##           5           5           5           5           5           5

AN0data <- data.frame(ID, Country, Days)
str(AN0data)

## 'data.frame':   30 obs. of  3 variables:
##  $ ID      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Country: Factor w/ 6 levels "France","Germany",...: 4 4 4 4 4 2 2 2 2 2 ...
##  $ Days   : num  4.3 7.5 14.78 6.35 8.01 ...

head(AN0data)

##   ID Country      Days
## 1  1 Norway  4.301959
## 2  2 Norway  7.504295
## 3  3 Norway 14.779288
## 4  4 Norway  6.346031
## 5  5 Norway  8.013474
## 6  6 Germany 5.556894

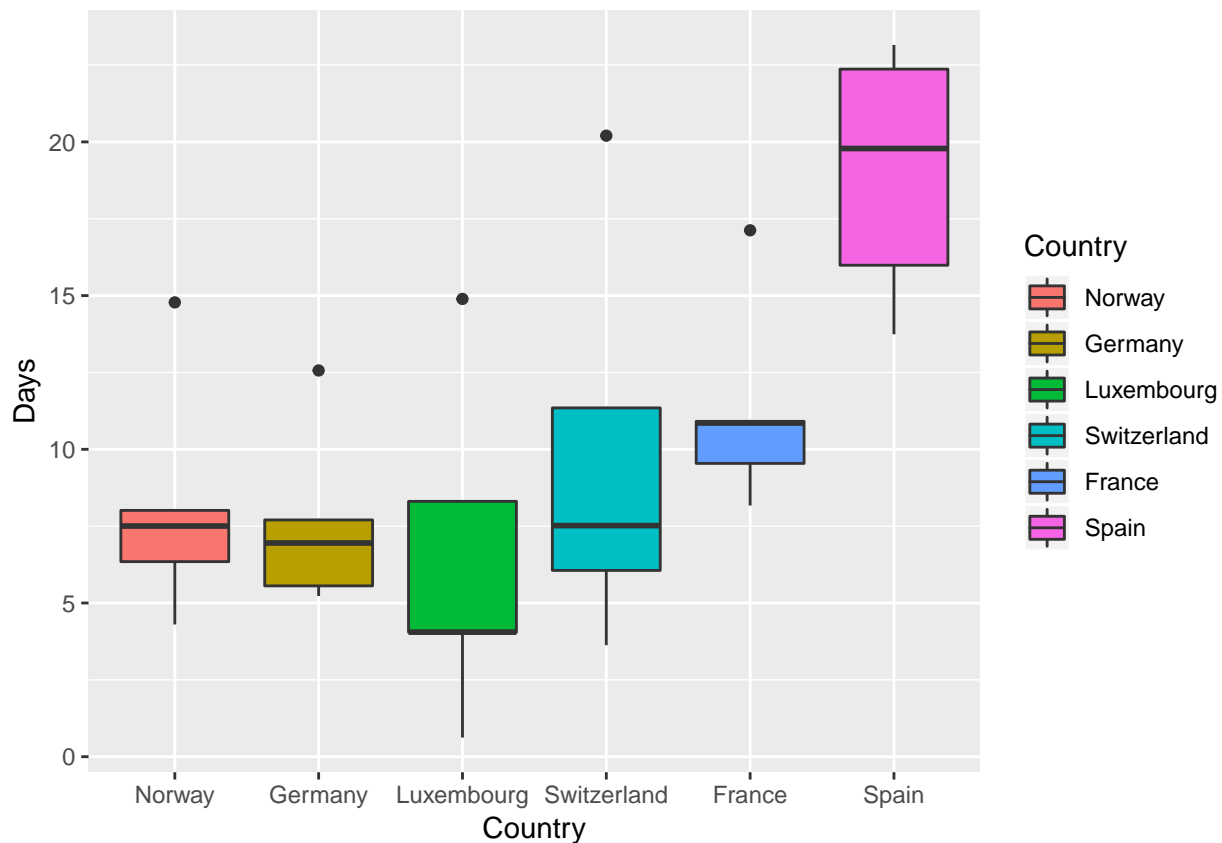
# Basic ANOVA model in R
AN0model <- aov(Days~Country, data=AN0data)
print(AN0model)

## Call:
##   aov(formula = Days ~ Country, data = AN0data)
##
## Terms:
##               Country Residuals
## Sum of Squares  521.0122  498.4868
## Deg. of Freedom      5       24
##
## Residual standard error: 4.557443
## Estimated effects may be unbalanced

print(summary(AN0model))

##              Df Sum Sq Mean Sq F value    Pr(>F)
## Country        5  521.0   104.20    5.017 0.00276 **
## Residuals     24  498.5    20.77
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

AN0data$Country = with(AN0data, reorder(Country, ID))
ANOPlot <- ggplot(data=AN0data,
                  aes(x=Country, y=Days,
                      fill= Country)) +
  geom_boxplot()
print(ANOPlot)
```



Halving the sample size inconsistently showed a p values of less than 0.05. I'll try a sample size of 6.

#Data frame construction for ANOVA

```
nGroup <- 6
nName <- c("Norway", "Germany", "Luxembourg", "Switzerland", "France", "Spain")
nSize <- c(6, 6, 6, 6, 6, 6) #changed sample size from 10 to 5 plants per country.
nMean <- c(3, 6, 9, 12, 15, 18)
nSD <- c(5, 5, 5, 5, 5, 5)
ID <- 1:(sum(nSize))
Days <- c(rnorm(n=nSize[1], mean=nMean[1], sd=nSD[1]),
          rnorm(n=nSize[2], mean=nMean[2], sd=nSD[2]),
          rnorm(n=nSize[3], mean=nMean[3], sd=nSD[3]),
          rnorm(n=nSize[4], mean=nMean[4], sd=nSD[4]),
          rnorm(n=nSize[5], mean=nMean[5], sd=nSD[5]),
          rnorm(n=nSize[6], mean=nMean[6], sd=nSD[6]))
Country <- rep(nName, nSize)
table(Country)
```

```
## Country
##      France      Germany  Luxembourg      Norway      Spain Switzerland
##          6          6          6          6          6          6
```

```
AN0data <- data.frame(ID, Country, Days)
str(AN0data)
```

```
## 'data.frame':   36 obs. of  3 variables:
##  $ ID      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Country: Factor w/ 6 levels "France","Germany",...: 4 4 4 4 4 4 2 2 2 2 ...
##  $ Days   : num  6.178 0.721 -7.324 0.232 10.068 ...
```

```
head(ANodata)
```

```
##   ID Country      Days
## 1  1  Norway  6.1778632
## 2  2  Norway  0.7214825
## 3  3  Norway -7.3237884
## 4  4  Norway  0.2317637
## 5  5  Norway 10.0683435
## 6  6  Norway  8.7785222
```

```
# Basic ANOVA model in R
```

```
ANOmodel <- aov(Days~Country, data=ANodata)
print(ANOmodel)
```

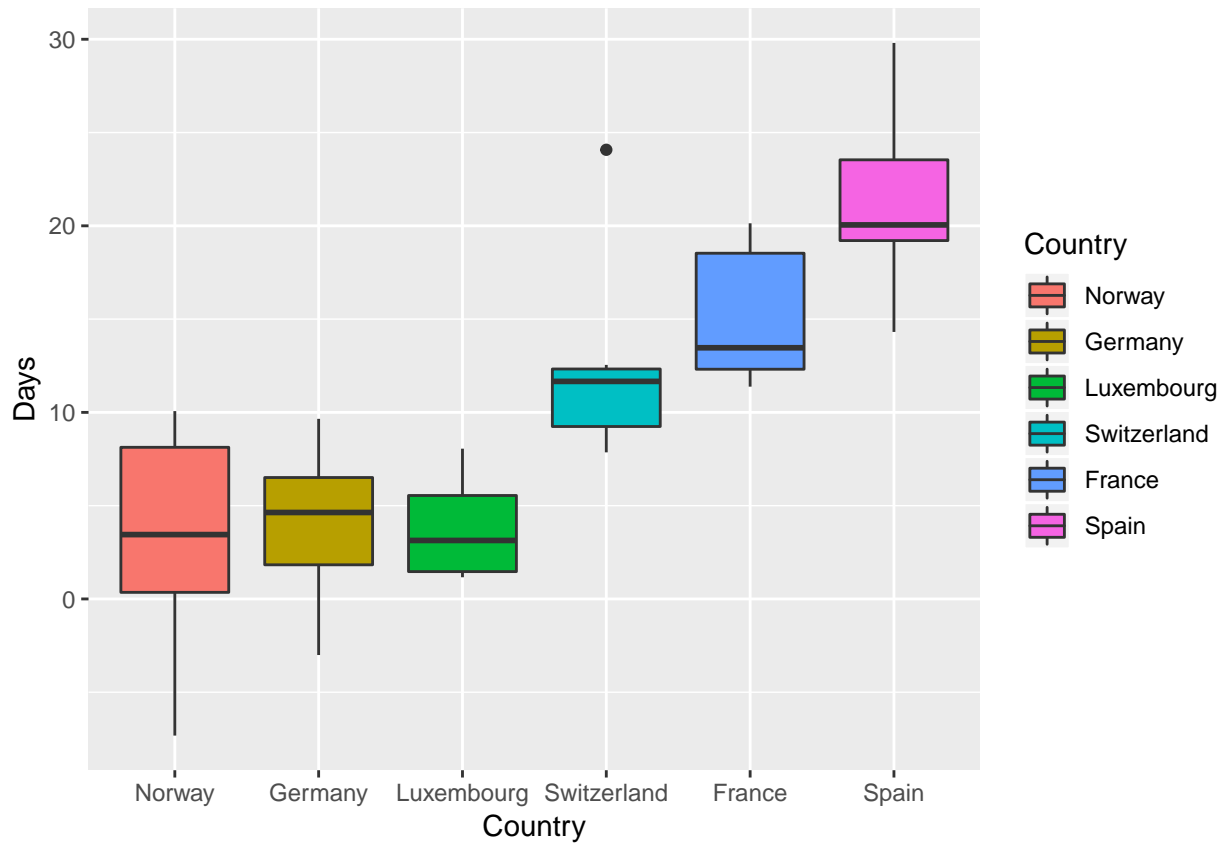
```
## Call:
##   aov(formula = Days ~ Country, data = ANodata)
##
## Terms:
##               Country Residuals
## Sum of Squares 1699.9934  742.0777
## Deg. of Freedom      5         30
##
## Residual standard error: 4.973522
## Estimated effects may be unbalanced
```

```
print(summary(ANOmodel))
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Country         5 1700.0   340.0   13.74 5.18e-07 ***
## Residuals      30  742.1    24.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ANodata$Country = with(ANodata, reorder(Country, ID))
```

```
ANOplot <- ggplot(data=ANodata,
                  aes(x=Country, y=Days,
                      fill= Country)) +
  geom_boxplot()
print(ANOplot)
```



After multiple runs of the code, the p value remains below 0.05. Therefore, the minimum sample size I would need in order to detect a statistically significant effect is 6.

Conclusion: Base on this exercise, I would only accept my p value if there is a sample size of no less than 6 centaurea plants per region and a difference of no less than 3 days between each mean.