Homework#7

cuden

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" exerciese, 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 folling order of flowering dates would be predicted for each popualtion:

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

Sinse 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 flowing 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, and 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, luxerberg'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 samples we will have in each group. "rule of ten:" at lea
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 treatme

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 m
```

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

```
## '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  7.01 5.65 7 14.59 3.46 ...
```

```
head(ANOdata)
```

```
##   ID Country      Days
## 1  1  Norway  7.007915
## 2  2  Norway  5.647440
## 3  3  Norway  6.996884
## 4  4  Norway 14.589500
## 5  5  Norway  3.460492
## 6  6  Norway  3.125660
```

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
ANOmodel <- aov(Days~Country, data=ANOdata) #"analysis of variance" response is function of (~) predict
print(ANOmodel)
```

```
## Call:
##    aov(formula = Days ~ Country, data = ANOdata)
##
## Terms:
##                   Country Residuals
## Sum of Squares  7996.595  1265.976
## Deg. of Freedom        5        54
##
## Residual standard error: 4.8419
## Estimated effects may be unbalanced
```

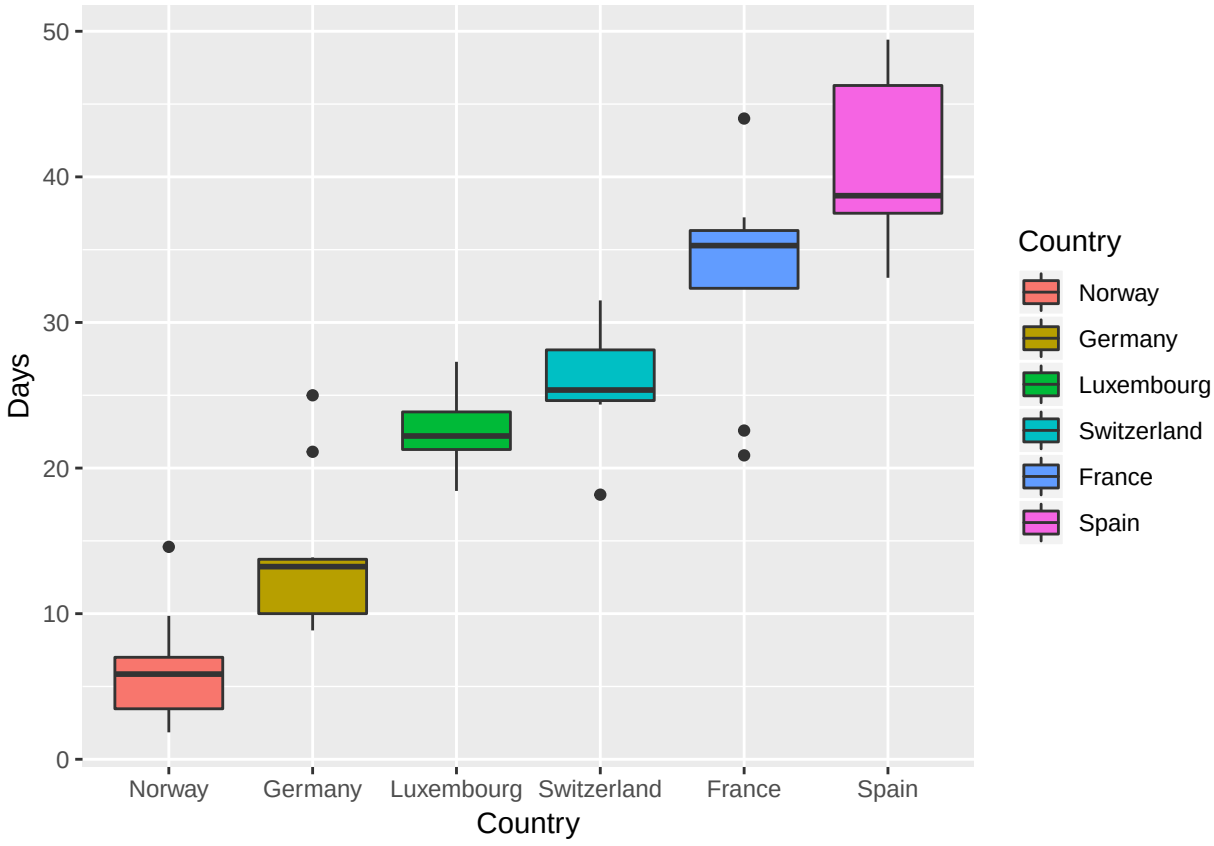
```
print(summary(ANOmodel))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Country      5   7997  1599.3   68.22 <2e-16 ***
## Residuals   54   1266    23.4
```

```
## ---
## 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)
```

```r
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
```

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

```
## '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  2.91 1.94 13.97 5.25 10.26 ...
```

```r
head(ANOdata)
```

```
##   ID Country      Days
## 1  1  Norway  2.907755
## 2  2  Norway  1.940979
## 3  3  Norway 13.970286
## 4  4  Norway  5.250135
## 5  5  Norway 10.255766
## 6  6  Norway  3.774189
```

```r
# Basic ANOVA model in R
ANOmodel <- aov(Days~Country, data=ANOdata)
print(ANOmodel)
```

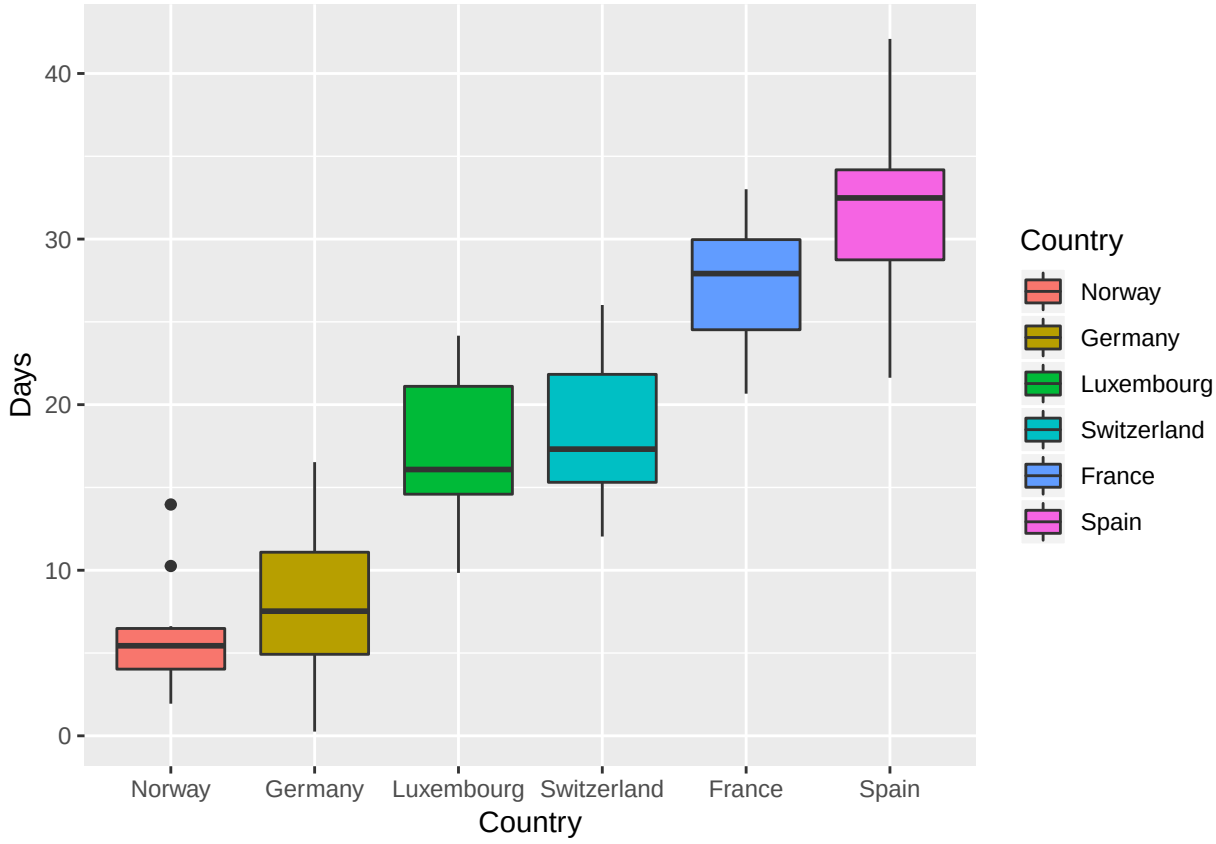
```
## Call:
##    aov(formula = Days ~ Country, data = ANOdata)
##
## Terms:
##                   Country Residuals
## Sum of Squares  5033.958  1185.087
## Deg. of Freedom         5        54
##
## Residual standard error: 4.684661
## Estimated effects may be unbalanced
```

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

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Country      5   5034  1006.8   45.88 <2e-16 ***
## Residuals   54   1185    21.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
ANOdata$Country = with(ANOdata, reorder(Country, ID))
ANOPlot <- ggplot(data=ANOdata,
```

```
                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
```

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

```
## '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  -1.457 2.293 0.337 2.692 -1.374 ...
```

```r
head(ANOdata)
```

```
##   ID Country       Days
## 1  1  Norway -1.4566205
## 2  2  Norway  2.2925833
## 3  3  Norway  0.3368022
## 4  4  Norway  2.6918676
## 5  5  Norway -1.3737293
## 6  6  Norway  5.0771219
```

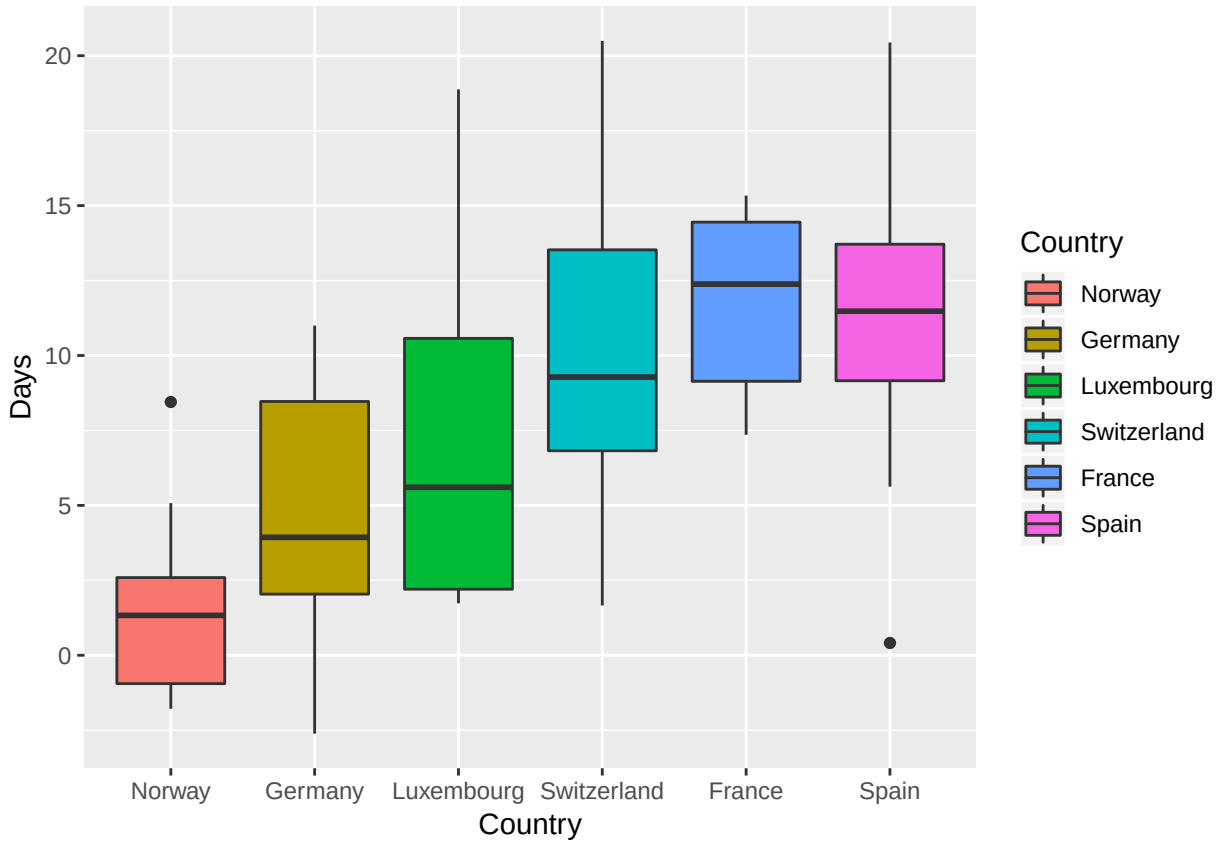
```r
# Basic ANOVA model in R
ANOmodel <- aov(Days~Country, data=ANOdata)
print(ANOmodel)
```

```
## Call:
##    aov(formula = Days ~ Country, data = ANOdata)
##
## Terms:
##                   Country Residuals
## Sum of Squares   823.4726 1223.2045
## Deg. of Freedom         5        54
##
## Residual standard error: 4.759405
## Estimated effects may be unbalanced
```

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

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Country      5  823.5  164.69   7.271 2.87e-05 ***
## Residuals   54 1223.2   22.65
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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



-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
```

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

```
## '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  1.04 1.81 -2.85 2.47 2.65 ...
```

```
head(ANOdata)
```

```
##   ID Country      Days
## 1  1  Norway  1.037413
## 2  2  Norway  1.811985
## 3  3  Norway -2.848036
## 4  4  Norway  2.473329
## 5  5  Norway  2.647493
## 6  6  Norway 11.451506
```

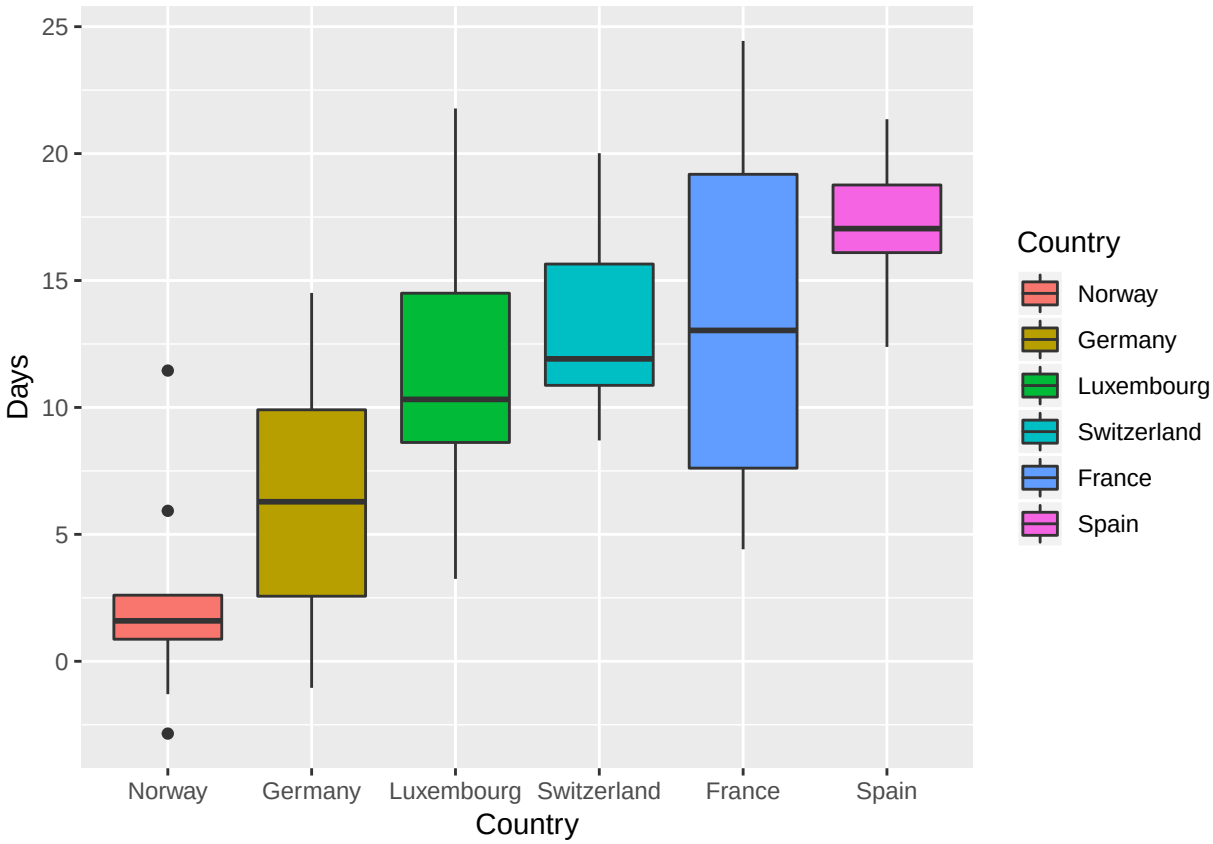
```
# Basic ANOVA model in R
ANOmodel <- aov(Days~Country, data=ANOdata)
print(ANOmodel)
```

```
## Call:
##    aov(formula = Days ~ Country, data = ANOdata)
##
## Terms:
##                   Country Residuals
## Sum of Squares  1459.601  1279.373
## Deg. of Freedom        5        54
##
## Residual standard error: 4.867453
## Estimated effects may be unbalanced
```

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

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Country      5   1460  291.92   12.32 5.46e-08 ***
## Residuals   54   1279   23.69
## ---
## 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 signficant relationship between country and flowering time unless the differene 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 smaple size form 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
```

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

```
## '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  8.433 -1.696 0.619 4.336 -4.014 ...
```

```
head(ANOdata)
```

```
##   ID Country       Days
## 1  1  Norway  8.4326324
## 2  2  Norway -1.6959214
## 3  3  Norway  0.6189368
## 4  4  Norway  4.3363102
## 5  5  Norway -4.0138094
## 6  6 Germany -2.0668979
```

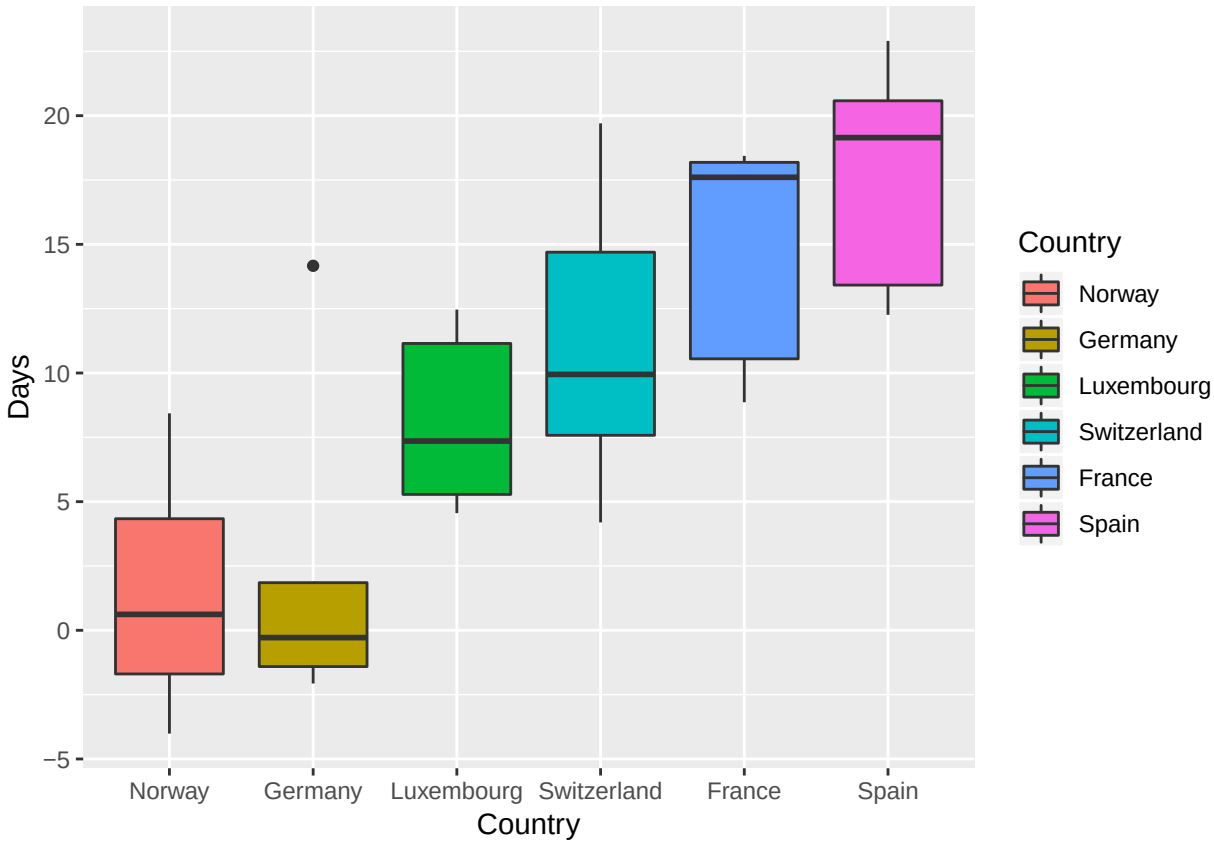
```
# Basic ANOVA model in R
ANOmodel <- aov(Days~Country, data=ANOdata)
print(ANOmodel)
```

```
## Call:
##    aov(formula = Days ~ Country, data = ANOdata)
##
## Terms:
##                   Country Residuals
## Sum of Squares  1057.933   646.841
## Deg. of Freedom        5        24
##
## Residual standard error: 5.191503
## Estimated effects may be unbalanced
```

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

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Country      5 1057.9  211.59   7.851 0.000169 ***
## Residuals   24  646.8   26.95
## ---
## 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)
```



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 smaple size form 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
```

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

```
## '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  1.76 -4.16 5.27 6.28 1.05 ...
```

```r
head(ANOdata)
```

```
##   ID Country      Days
## 1  1  Norway  1.764008
## 2  2  Norway -4.156448
## 3  3  Norway  5.271520
## 4  4  Norway  6.281011
## 5  5  Norway  1.045006
## 6  6  Norway  5.832870
```

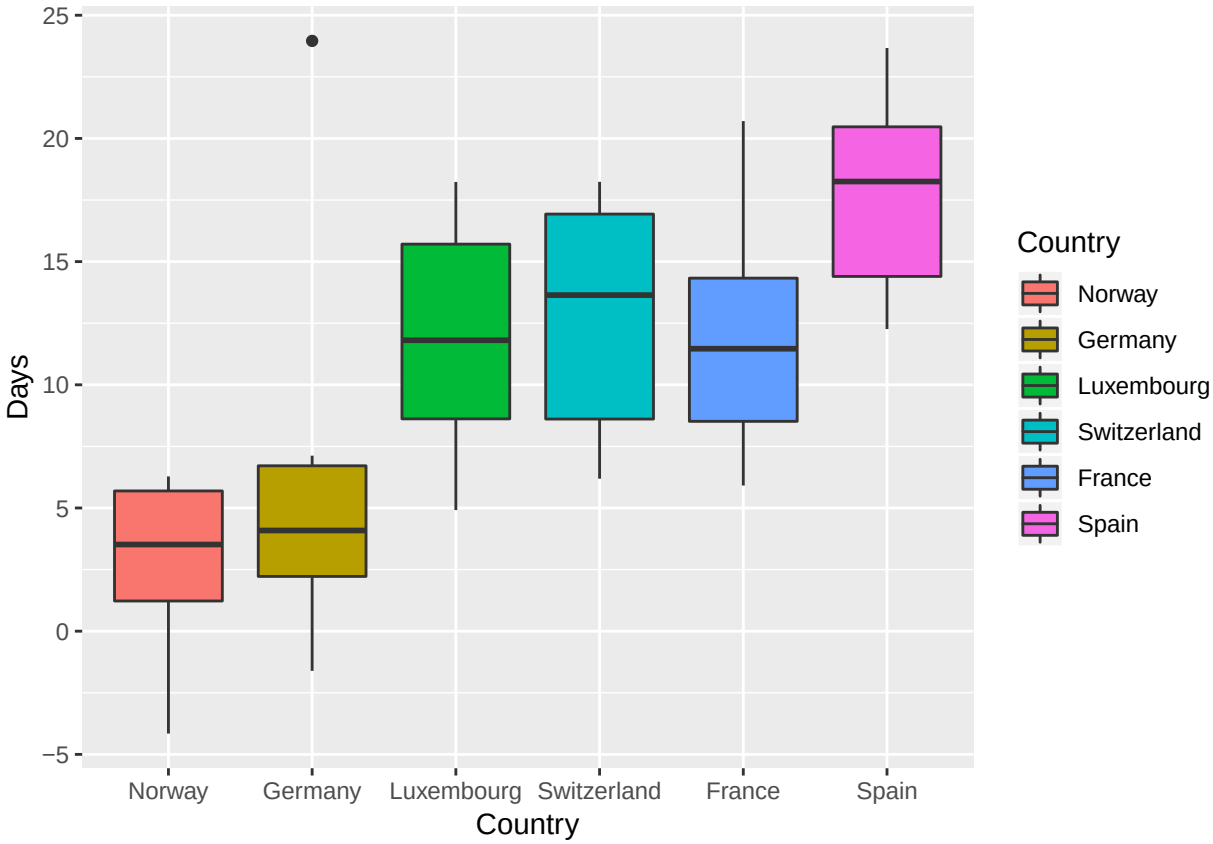
```r
# Basic ANOVA model in R
ANOmodel <- aov(Days~Country, data=ANOdata)
print(ANOmodel)
```

```
## Call:
##    aov(formula = Days ~ Country, data = ANOdata)
##
## Terms:
##                   Country Residuals
## Sum of Squares  833.7895  983.4232
## Deg. of Freedom        5        30
##
## Residual standard error: 5.72545
## Estimated effects may be unbalanced
```

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

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Country      5  833.8  166.76   5.087 0.00171 **
## Residuals   30  983.4   32.78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
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 diffference of no less than 3 days between each mean.