# INSH5301 Intro Computational Statistics

*Ali Banijamali*

*02/24/2020*

**1. Write a function that checks if a number is a prime. The output of the function should be a logical value, TRUE if the number is prime and FALSE if not. A prime number is a positive integer greater than 1 that is divisible without remainder only by 1 and itself. Then, generate a 10 by 10 matrix of random integers and count how many prime numbers are in it.**

```
is.prime.no <- function(number){
  # The function returns True, if the number is prime
  if (number == 2){ # Evaluate the case where the number is 2 separately
    T
  } else if (any(number %% 2:round(number/2) == 0)){
    F
  } else {T}
}

# Checking number 1:
is.prime.no(1)
```

```
## [1] FALSE
```

```
# Checking number 2:
is.prime.no(2)
```

```
## [1] TRUE
```

```
# Checking number 101:
is.prime.no(101)
```

```
## [1] TRUE
```

```
# Populating a 10X10 matrix of random integers:
# Generate 100 integer numbers between 1 to 50
ten.by.ten.mat <- matrix(sample.int(50, 100, replace=T), nrow=10, ncol=10)
ten.by.ten.mat
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]    1   16   25   41   49   40   29   49   21    38
##  [2,]   39    1   12   30   10   12   17   10   14    30
##  [3,]   37    6   47   31   14   22   32   38   29    43
##  [4,]   30   19   41   49   47    9   11   21   33    23
##  [5,]    5   31   23    7   47   48    3   12    1    16
##  [6,]    1   19   11    8   21   43   47    6    5     2
##  [7,]   42    8   20   33    7   16    5   50   23    21
##  [8,]    3   12   17    3    5   20    9   32   28    42
##  [9,]    3   48   11   43   38   37   17   22   11     8
## [10,]   34   35    3    5   19   47   30   36   50    12
```

```
# Counting the prime numbers in the matrix:
results <- lapply(ten.by.ten.mat, is.prime.no)
```

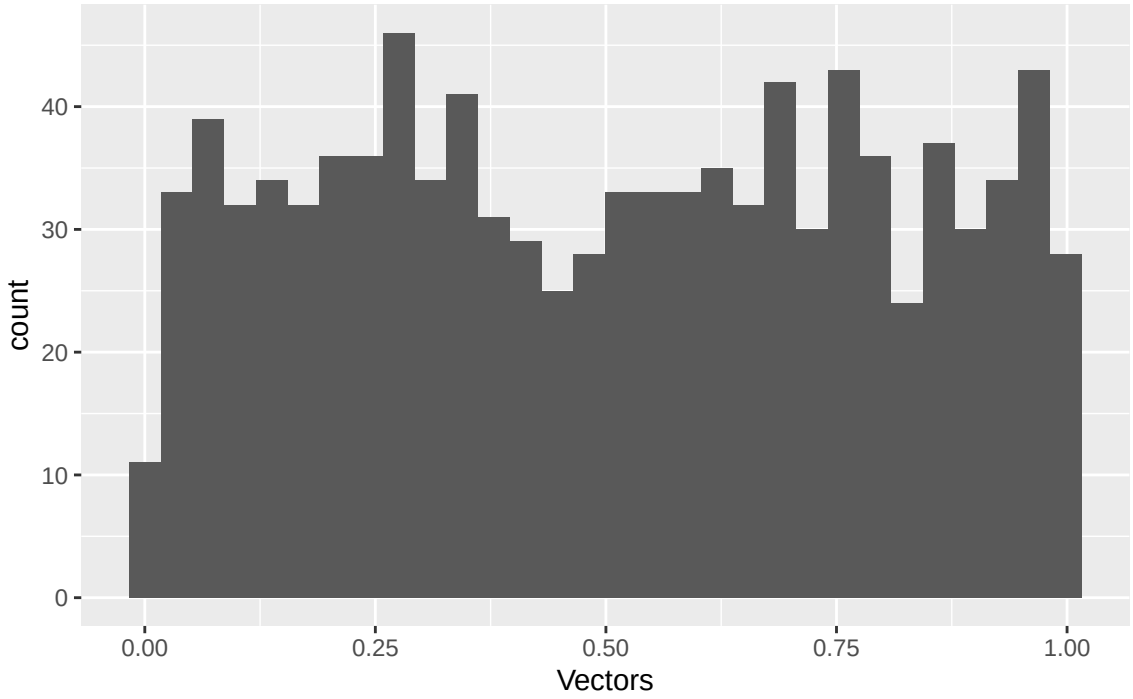
```r
length(results[results==T])
```

```
## [1] 42
```

**2. Write a function that given an integer k and a sample size n generates k random vectors using the uniform distribution with bounds (0, 1) of length n and add them. Using this function, produce 5 vectors, each with sample size n = 1,000, but with different values of k; in particular, use (k = 1, k = 2, k = 3, k = 4, and, k = 5). Make a histogram using ggplot2 of each vector.**

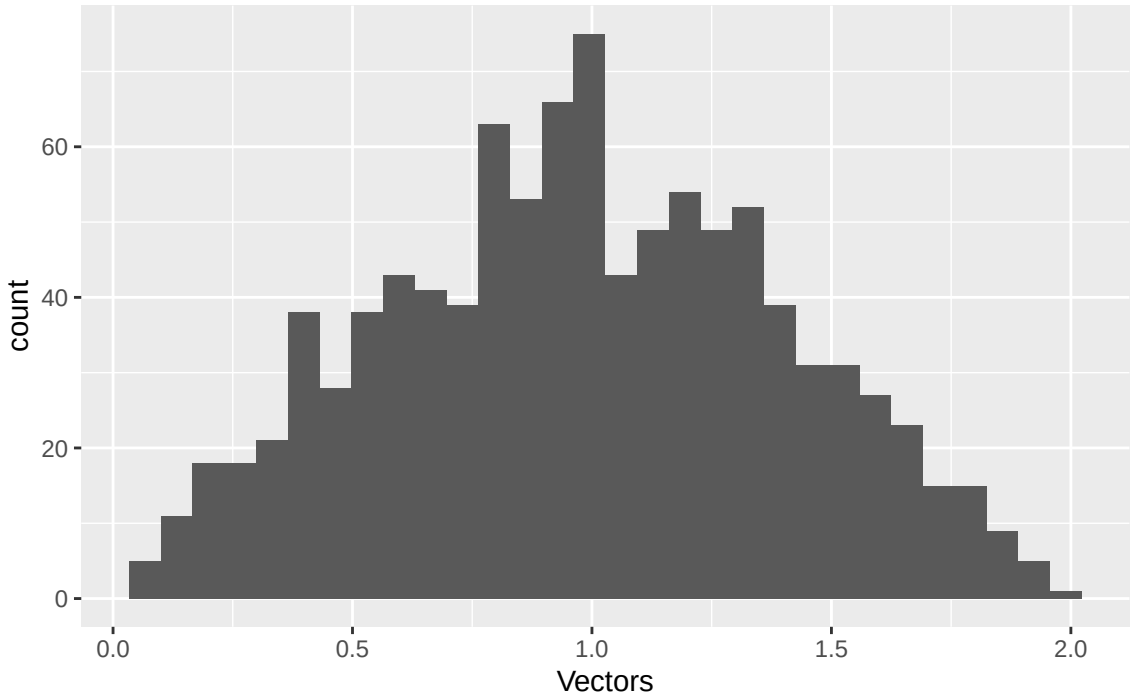
```r
library(ggplot2)

custom.func <- function(k, n){
  # k is an integer, representing the number of random vectors
  # n is the sample size of each vector
  if(k==1){
    vector <- runif(n, 0, 1) # uniform dist of size n, min=0, max=1
    return(vector)
  }else{
    vector <- runif(n, 0, 1)
    for(i in 2:k){
      vector <- vector + runif(n, 0, 1)
    }
    return(vector)
  }
}

# Making the histograms:
plots <- list()
for(i in 1:5){ # for 5 k values
  # Making the vector:
  v_i <- data.frame(vec = custom.func(k=i, n=1000))

  # Saving the histogram:
  plots[[i]] <- ggplot(v_i, aes(x=vec)) + geom_histogram() +
    ggtitle(paste("for k = ", i, ":", sep="")) + xlab("Vectors")
}

# Plotting the histograms:
plots[[1]]
```

for k = 1:



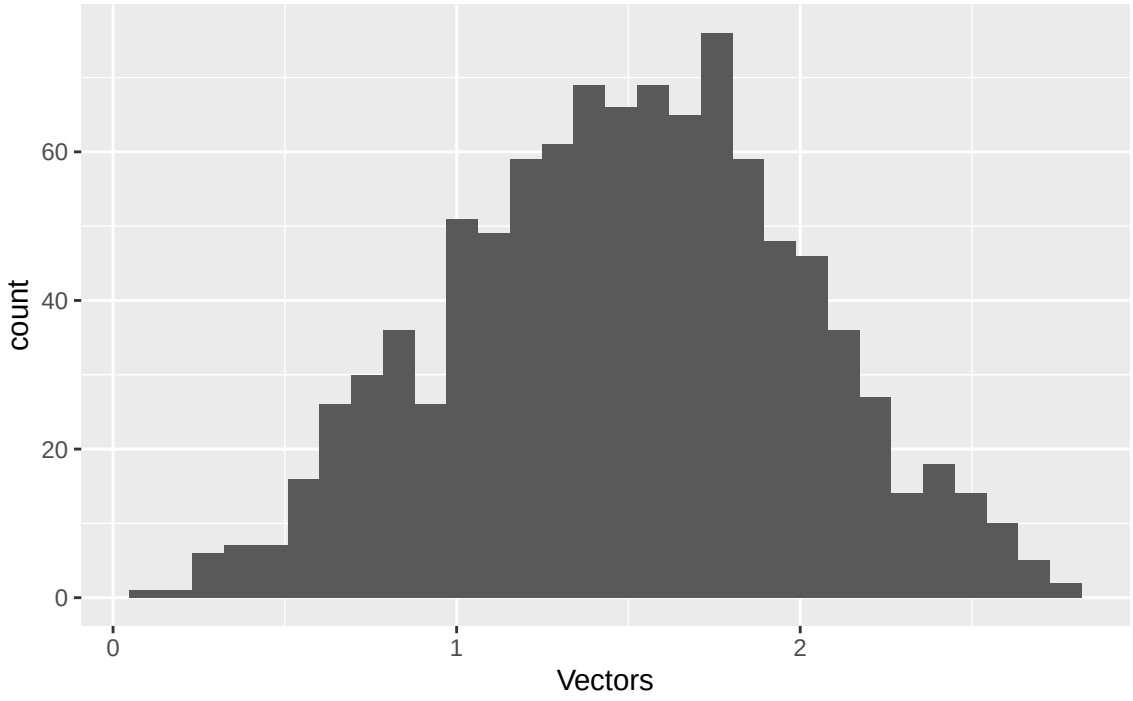

```
plots[[2]]
```

for k = 2:



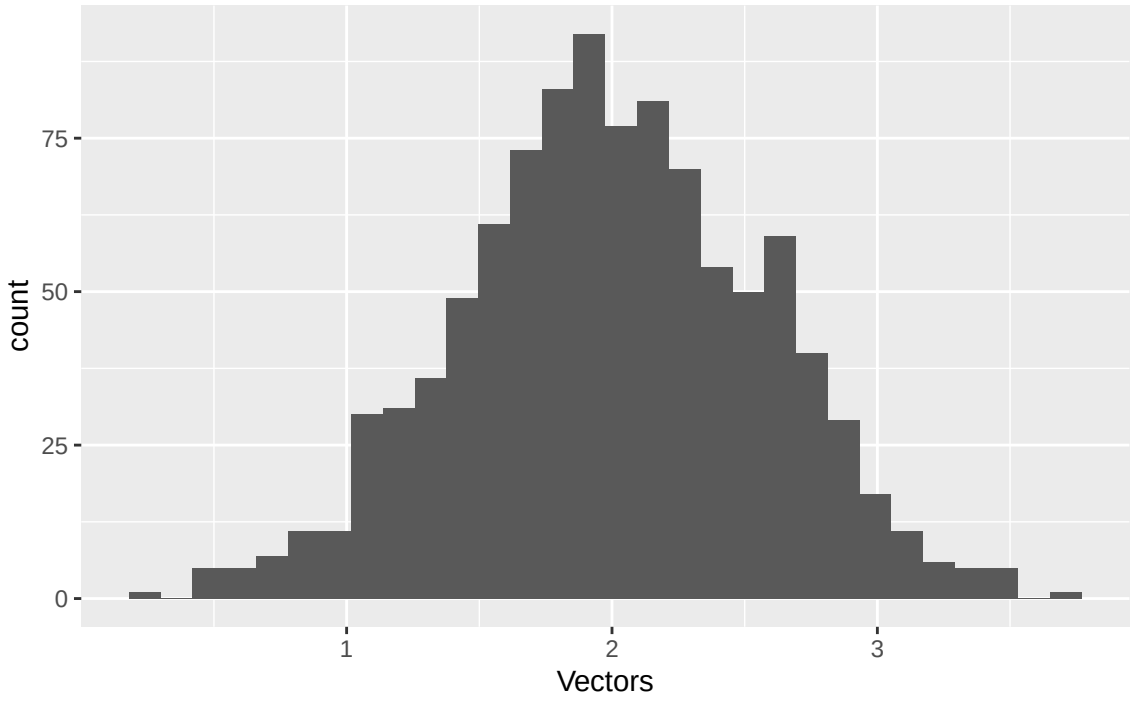

```
plots[[3]]
```

for k = 3:

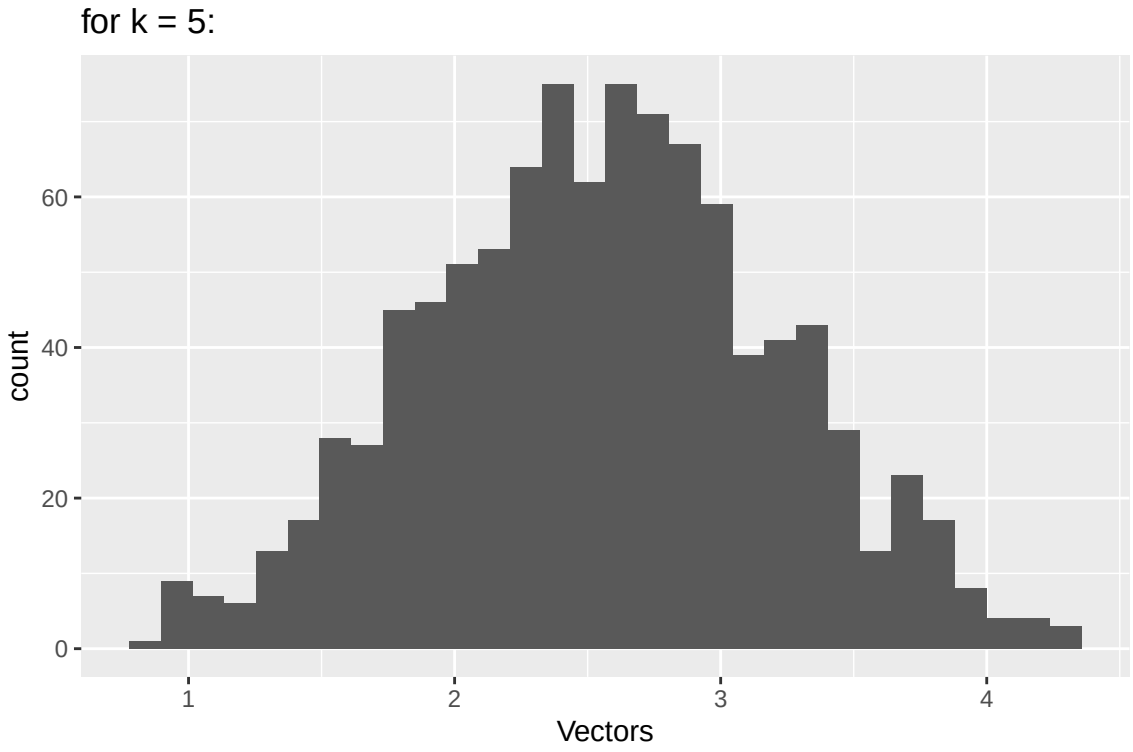



```
plots[[4]]
```

for k = 4:




```
plots[[5]]
```




### 3.a. You roll two 6-sided dies twice and add the result. a. What's the probability of getting a 12?

ANS.
The only possible case to have a sum of 12 is to have a 6 on the first die and another 6 on the second die. This translates to:
$P_{6\ on\ die\ 1} \times P_{6\ on\ die\ 2} = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$

### 3.b. What's the probability of getting a result less or equal to 6?

ANS.
Total number of possible outcomes is $6 \times 6 = 36$
Our desired outcomes are:
(1,1), (1,2), (1,3), (1,4), (1,5),
(2,1), (2,2), (2,3), (2,4),
(3,1), (3,2), (3,3),
(4,1), (4,2),
(5,1)
The count of our desired outcomes is: 15
Therefore the probability will be: $P_{less\ than\ and\ equal\ to\ 6} = \frac{15}{36}$

### 4. The average number of snow days in Boston in a year is 22.

### 4.a. Assuming snow days follow a Poisson distribution, calculate (using R) the probability of getting 10 or less days of snow in a year.

ANS. In our Poisson's distribution, mean is 22 (Average rainy days per year). We want the probability of 10 or less rainy days in Boston. This means:
$P_{10\ or\ less\ rainy\ days} = P_{1\ rainy\ days} + P_{2\ rainy\ days} + ... + P_{10\ rainy\ days}$ (+: OR)

```r
# The probability of event x with Poisson's distribution where the mean is lambda, is:
# dpois(x, lambda), we must sum 10 of these for each day:
sum(dpois(1:10, 22))
```

```
## [1] 0.0035466
```

**4.b. Knowing that in February the probability of a snow day is 20 percent. If it's snowing in Boston, what's the probability that the month is February? (5 pt)**

ANS.
This problem should be solved based on the Bayesian inference:
Since it is stated that it is snowing, therefore $P_{snow} = 1$
$P_{February/snow} = \frac{P_{February} \times P_{snow/February}}{P_{snow}} = \frac{\frac{1}{12} \times 0.2}{1} = 0.01666667$

**5. You want to know how many hours of sleep the average college student gets. You survey 10 stuents and get the following data (in hours): 7,6,5,8,6,6,4,5,8,12. Conduct a one-sided statistical test to test if the population mean is less or equal to 7.**

ANS.
First let's define our null and research hypothesis:
We are only interested in $\mu \leq 7$, so our Null hypothesis is:
$H_0 : \mu > 7$
and the research/alternative hypothesisis:
$H_a : \mu \leq 7$
Now let's do the one-sided test:

```r
# Student data:
sleep.hours <- c(7,6,5,8,6,6,4,5,8,12)

t.test(sleep.hours, alternative="less", mu=7)
```

```
##
##  One Sample t-test
##
## data:  sleep.hours
## t = -0.41917, df = 9, p-value = 0.3425
## alternative hypothesis: true mean is less than 7
## 95 percent confidence interval:
##      -Inf 8.011953
## sample estimates:
## mean of x
##       6.7
```

As can be seen, the t-statistics is -1.8164, the degree of freedom is (10-1=9), and the p-value is 0.05134. Although close, but p-value is above 0.05. Based on a 5 percent p-value we CAN NOT reject the null hypothesis and therefore our research hypothesis can't be proven.

MANUAL:
$\overline{x} = \frac{7+6+5+8+6+6+4+5+8+12}{10} = 6.7$
$SD = \sqrt{\frac{(7-6.7)^2+(6-6.7)^2+(5-6.7)^2+(8-6.7)^2+(6-6.7)^2+(6-6.7)^2+(4-6.7)^2+(5-6.7)^2+(8-6.7)^2+(12-6.7)^2}{10-1}} = 2.263$
$\mu = 7$

$se = \frac{SD}{\sqrt{n}} = \frac{2.263}{3.163} = 0.715$ $df = 10 - 1 = 9$
$T_{statistics} = \frac{\bar{x}-\mu}{se} = \frac{6.7-7}{0.715} = -0.42$

Threshold for one-tail based on 0.05 p-value:

```
qt(0.05, 10-1)
```

```
## [1] -1.833113
```

Since T-statistic (-0.42), is above -1.83, we CAN NOT reject the null hypothesis and therefore the alternative hypothesis is not proven.

## 6. You survey the same 10 people during finals period, and get the following hours: 5,4,5,7,5,4,5,4,6,12. Do college students get significantly less sleep than usual during finals?

ANS.
Now we have a different case. We would like to compare two means. Let's define our Null and Alternative hypothesis:
$H_0 : \mu_{difference\ in\ sleeping\ hours\ (normal-exam\ night)} = 0$
$H_a : \mu_{difference\ in\ sleeping\ hours\ (normal-exam\ night)} > 0$ : Meaning that students are sleeping significantly more hours on regular nights when they don't have exams.

Now let's do the test. Again I am using a one-sided t-test again, but I first have to subtract the sleeping hours of exam nights from normal nights, then check if these values are significantly greater than zero, meaning students sleep significantly more on regular nights:

```
# Student data:
no.exam.sleep.hours <- c(7,6,5,8,6,6,4,5,8,12)
exam.sleep.hours    <- c(5,4,5,7,5,4,5,4,6,12)
sleep.difference    <- no.exam.sleep.hours-exam.sleep.hours

# Now let's take the test:
t.test(sleep.difference, alternative="greater", mu=0)
```

```
##
##  One Sample t-test
##
## data:  sleep.difference
## t = 3, df = 9, p-value = 0.007478
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  0.3889624       Inf
## sample estimates:
## mean of x
##         1
```

As can be seen, based on the p-value of 0.007478, we CAN reject the null hypothesis, meaning that the sleeping hours during final night ARE significantly less than normal nights. in other words, student sleep more on regular nights.

MANUAL:

```
# Student data:
no.exam.sleep.hours <- c(7,6,5,8,6,6,4,5,8,12)
exam.sleep.hours    <- c(5,4,5,7,5,4,5,4,6,12)
sleep.difference    <- no.exam.sleep.hours-exam.sleep.hours
```

```r
# Difference in sleeping hours:
sleep.difference
```

```
##  [1]  2  2  0  1  1  2 -1  1  2  0
```

```r
# Mean:
mean(sleep.difference)
```

```
## [1] 1
```

```r
# Standard Deviation:
sd(sleep.difference)
```

```
## [1] 1.054093
```

$se = \frac{SD}{\sqrt{n}} = \frac{1.054093}{\sqrt{10}} = 0.3333335$
$df = n - 1 = 9$
$T - statistic = \frac{\overline{x} - \mu}{se} = \frac{1-0}{0.3333335} = 2.999999$
Based on 0.05 p-value, lower and upper bounds are:

```r
# Lower bound:
qt(0.025,9)
```

```
## [1] -2.262157
```

```r
# Upper bound:
qt(0.975,9)
```

```
## [1] 2.262157
```

Since 2.99 is outside of this region [-2.262157, 2.262157], we CAN reject the null hypothesis and the alternative hypothesis is proved.

**7. You are a very bad gardener, and hypothesize that feeding houseplants with caffeine might help them grow better. You perform an experiment to test your hypothesis. To three separated groups of plants you gave them:**

**(1) water spiked with diet coke,**

**(2) water spiked with coffee, and,**

**(3) water alone.**

**The table below summarize the result for each group.**

| Condition | Mean.Days.Alive_Life.Expectancy | SD | n |
|---|---|---|---|
| Water | 50 | 10 | 20 |
| Diet Coke | 42 | 7 | 15 |
| Coffee | 52 | 8 | 10 |

**Test if there is any statistically significant difference among the groups' life expectancy.**

ANS.
Here, I'd like to define the Null and Alternative hypothesis as follows, since we want to show that these are actually different:
$H_0 : \mu_{water} = \mu_{diet\ coke} = \mu_{coffee}$

$H_a$ : *At least one of these methods is different*
I am going to use R to get an F test:

```
set.seed(1)
# 3 vectors with foods and respective mean and sd:
water <- data.frame(Party=as.factor("Water"), Age=rnorm(20, 50, 10))
coke <- data.frame(Party=as.factor("Coke"), Age=rnorm(15, 42, 7))
coffee <- data.frame(Party=as.factor("Coffee"), Age=rnorm(10, 52, 8))

food.life <- rbind(water, coke, coffee)

anova <- aov(food.life[,2]~food.life[,1], data=food.life)
summary(anova)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## food.life[, 1]  2   1166   582.9   10.21 0.000244 ***
## Residuals      42   2398    57.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As can be seen the p-value (0.000244) is a lot smaller than 0.05, therefore the null hypothesis IS rejected and the alternative hypothesis is proved. This means that there is difference between life expectancy of plant under different feeding methods.
Our $df_1 = 2$ and $df_2 = 42$

MANUAL:
F-statistics $= \frac{Average\ \ variance\ \ between\ \ groups}{Average\ \ variance\ \ within\ \ groups}$
Between Variance $= \frac{n_1(\bar{y_1}-\bar{y})^2+...+n_G(\bar{y_G}-\bar{y})^2}{G-1}$
and,
Within Variance $= \frac{(n_1-1)s_1^2+...+(n_G-1)s_G^2}{N-G}$
Overall mean $\bar{y} = \frac{20\times50+15\times42+10\times52}{20+15+10} = \frac{2150}{45} = 47.77778 \approx 47.78$

BV $= = \frac{20(50-47.78)^2+15(42-47.78)^2+10(52-47.78)^2}{3-1} = 388.89$
WV $= = \frac{(20-1)\times10^2+(15-1)\times7^2+(10-1)\times8^2}{45-3} = 75.28$
F-statistic $= \frac{388.89}{75.28} = 5.166$

Degrees of freedom are:
$df_1 = 3 - 1 = 2$
$df_2 = 45 - 3 = 42$

And finally finding the F threshold:

```
qf(0.95,2,42)
```

```
## [1] 3.219942
```

The F-statistic is 5.166 well above 3.219942, therefore we CAN reject the null hypothesis and the alternative hypothesis is proven.

**8. You are a interested in knowing if knowing more than one language is related to memory. To test your hypothesis you run an experiment with 100 individuals, to each participant you asked if they speak at least two languagues and gave them a memory test. Using the test results you categorize each participant in a different groups: good memory, normal memory, and bad memory.**

| | Monolingual | At.Least.Bilingual |
|---|---|---|
| Good | 24 | 6 |
| Average | 40 | 5 |
| Bad | 23 | 2 |

### Test if the number of languages an individual can speak is related to memory.

ANS.
The best method is to do the Chi-squared test. First let's define the null and alternative hypothesis:

$H_0$ : *Memory and number of spoken languages are not related*
$H_a$ : *Memory and number of spoken languages are dependent*

```
memory.vs.language <- data.frame("monolingual"=c(24, 40, 23),
                                 "bilingual"=c(6, 5, 2),
                                 row.names=c("Good","Average","Bad"))

chisq.test(memory.vs.language)
```

```
##
##  Pearson's Chi-squared test
##
## data:  memory.vs.language
## X-squared = 1.9943, df = 2, p-value = 0.3689
```

Based on the p-value of 0.3689, we CAN NOT say that there is a relationship between memory and spoken languages, meaning that the null hypothesis WAS NOT rejected.

MANUAL:
First let's calculate the probabilities of individual variables:
$P_{Good} = \frac{24+6}{100} = 0.3$
$P_{Average} = \frac{40+5}{100} = 0.45$
$P_{Bad} = \frac{23+2}{100} = 0.23$
$P_{Monolingual} = \frac{24+40+23}{100} = 0.87$
$P_{At\ least\ Bilingual} = \frac{6+5+2}{100} = 0.13$

Next we can calculate the expected probabilities and $f_e$s for cells:

$P_{expected\ (mono\&good)} = P_{mono} \times P_{good} = 0.87 \times 0.3 = 0.261$
$f_e = P_{expected} \times total = 0.261 \times 100 = 26.1$

$f_{e\ mono\&ave} = P_{mono\&ave} \times 100 = 39.15$
$f_{e\ mono\&bad} = P_{mono\&bad} \times 100 = 20$

$f_{e\ bi\&good} = P_{bi\&good} \times 100 = 3.9$
$f_{e\ bi\&ave} = P_{bi\&ave} \times 100 = 5.8$
$f_{e\ bi\&bad} = P_{bi\&bad} \times 100 = 2.99$

$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e} \approx 3.549$

Now, the degree of freedom is:
$df = (r-1) \times (c-1) = 2 \times 1 = 2$

Based on this d.o.f and 95% threshold, we have:

```
qchisq(0.95, df=2)
```

```
## [1] 5.991465
```

Our chi-squared (3.549) is well below 5.991465, therefore we CAN NOT reject the null hypothesis. There is no relationship between memory and number of spoken languages.