Lab handout (Week 4: Sept. 24, 2019)

Consider the dataset SurveyFall2019.csv collected from the Getting To Know You Survey and use the following codes to load the dataset:

```
data <- read.csv("http://users.stat.umn.edu/~parky/SurveyFall2019.csv", header = T)
attach(data)
```

Let's explore some useful R functions.

a. Use dim() to identify the dimension of the dataset. What do the numbers in the output represent?

```
> dim(data)
[1] 358   26
```

358 is the number of observations, 26 is the total number of variables.

b. Use mean() to calculate the average idea.weight and average ideal.weight of students that exercise over 1 hour per day.

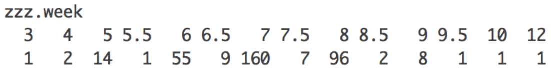
```
> mean(ideal.weight)
[1] 147.1416
> mean(ideal.weight[exercise>1])
[1] 152.9156
```

c. Use length() to find the number of female students and number of female students that exercise over 1 hour per day.

```
> length(which(gender=="Female"))
[1] 211
> length(which(gender=="Female" & exercise>1))
[1] 63
```

d. Obtain the frequency table of the zzz.week variable.

> table(zzz.week)



```
zzz.week
  3   4   5 5.5   6 6.5   7 7.5   8 8.5   9 9.5  10  12 
  1   2  14   1  55   9 160   7  96   2   8   1   1   1 
```

For question 1-2, we examine the distribution of the zzz.week variable (the number of hours slept per night during the week) for the STAT 3011 students. Use the frequency table to answer the following questions.

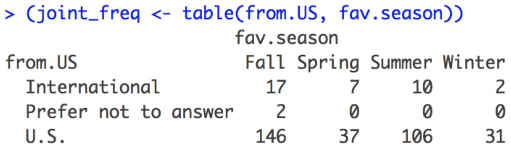
1. What is the probability that a randomly chosen STAT 3011 student sleeps less than 6 hours per night during the week?

$P(zzz.week < 6) = \frac{1+2+14+1}{358} = \frac{18}{358} = 0.050$

2. What is the probability that a randomly chosen STAT 3011 student sleeps 8 hours or more per night during the week?

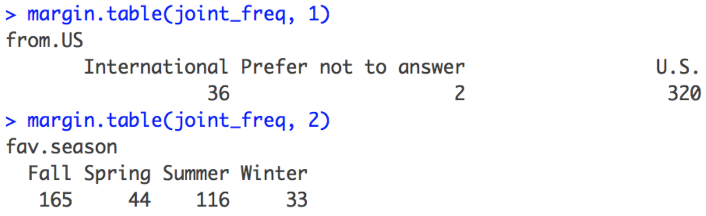
$P(zzz.week \geq 8) = \frac{96+2+8+1+1+1}{358} = \frac{109}{358} = 0.304$

e. Obtain the cross table for the distributions of the from.US and fav.season variables using the table(variable1,variable2) command. This command allows you to create a table for two variables at the same time. In statistics, we call this a cross-tabulation or two-way table.



```
> (joint_freq <- table(from.US, fav.season))
                      fav.season
from.US                Fall Spring Summer Winter
  International          17      7     10      2
  Prefer not to answer    2      0      0      0
  U.S.                  146     37    106     31
```

f. Obtain the marginal table for from.US and fav.season variables.



```
> margin.table(joint_freq, 1)
from.US
       International Prefer not to answer                 U.S.
                  36                    2                  320
> margin.table(joint_freq, 2)
fav.season
  Fall Spring Summer Winter
   165     44    116     33
```

For question 3-5, we examine the joint distribution of the from.US variable (whether the student is from US) and fav.season variable (the student's favorite season) for the STAT 3011 students. Use the frequency tables above to answer the following questions.

3. What is the probability that a randomly chosen STAT 3011 student's favorite season is winter?

$P(Winter) = \frac{33}{358} = 0.092$

4. What is the conditional probability that a randomly chosen STAT 3011 student's favorite season is winter, given that he/she is an international student?

$P(Winter \mid International) = \frac{P(Winter \cap International)}{P(International)} = \frac{2/358}{36/358} = 0.056$

5. Based on your answers to question 3 and 4, are the event that a student's favorite season is winter and the event that he/she is an international student independent? Explain.

Since $P(Winter) \neq P(Winter \mid International)$, the event that a student's favorite season is winter and the event that he/she is an international student are NOT independent.