Kursad Tosun October 22, 2019

# MATH 275 MIDTERM EXAM - Fall 2019

Answer the following questions using R Markdown. Answers without R codes (justification) will not be accepted. Upload your solution to Canvas by 10pm October 24, 2019. The file name should be in the form of "M275_MidtermExam_ Last Name followed by Initials.Rmd". For instance, M275_MidtermExam_TosunK.Rmd. Make your source code is well-documented and reproducible (i.e., I should be able to reproduce your results by running your submitted code as-is smoothly). Pay attention to good writing and communication of results. Include your name, course number (Math 275), and date in the header of your write-up.

**1. Install the `faraway` package.**

**2. Call the `faraway` package from library.**

**3.** Explore the `prostate` data.

```
data(prostate)
head(prostate)
? prostate
summary(prostate)
typeof(prostate$svi)
```

In this question you will export the `prostate` data into your computer with a different name. Before that, read the following.

To export a dataset e.g. `Math275Data` to a CSV file with a name `filename.csv`, use the `write.csv()` function.

```
write.csv(Math275Data, "filename.csv")         # Do not run this code.
```

This command creates a CSV file and saves it to your working directory, which by default is your home folder (for Mac and Linux users) or your 'My Documents' folder (for Windows users). To save the file somewhere other than in the working directory, enter the full path for the file.

```
write.csv(Math275Data, "~/whateverfolder/filename.csv", row.names=FALSE)        # for mac
write.csv(Math275Data, "C:/whateverfolder/filename.csv", row.names=FALSE)       # for windows
```

Now, you are ready for the question. **Export the `prostate` data from R to CSV file and save into your computer as `newprostatedata.csv`. Then, import the `newprostatedata.csv` file into R (as usual), and create a new object with the name `newpros`.**

**4.** Invasion of the muscular wall of the seminal vesicles by prostate cancer is generally regarded as a marker of poor prognosis (the cancer is harder to control) at the time of pathologic staging after radical prostatectomy. Seminal vesicle invasion (svi) is associated with increased risk of lymph node metastasis and tumor recurrence in patients with prostate cancer, and therefore knowledge of its presence at the time of diagnosis is an important factor for prognosis assessment and patient treatment.

To find the number of patients who had seminal vesicle invasion (`svi` is `1` in the `newpros`) we can use the `summary()` function.

```
summary(newpros$svi)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.2165  0.0000  1.0000
```

The output of `summary(newpros$svi)` shows that, the mean of `svi` is 0.2165. However, `svi` has only two values: `0` which corresponds to *no invasion* and `1` corresponds to *presence of invasion*. Actually `svi` is a categorical variable. Since R summarize `svi` with mean, it considers that `svi` has numeric values. Let's first check that.

```
typeof(newpros$svi)

## [1] "integer"
```

Now, let's make it categorical.

```
newpros$svi=factor(newpros$svi)
summary(newpros$svi)

##  0  1
## 76 21
```

We can also use the pipe operator `%>%` and `mutate()` function from `dplyr` package.

```
library(dplyr)
newpros <- newpros %>%
  mutate(svi=factor(svi))
```

We can even make our variable self explanatory by manupulating the dataset `newpros`.

```
newpros <- newpros %>%
  mutate(svi=factor(svi,levels = c("1", "0"),labels = c("yes", "no")))
summary(newpros$svi)

## yes  no
##  21  76
```

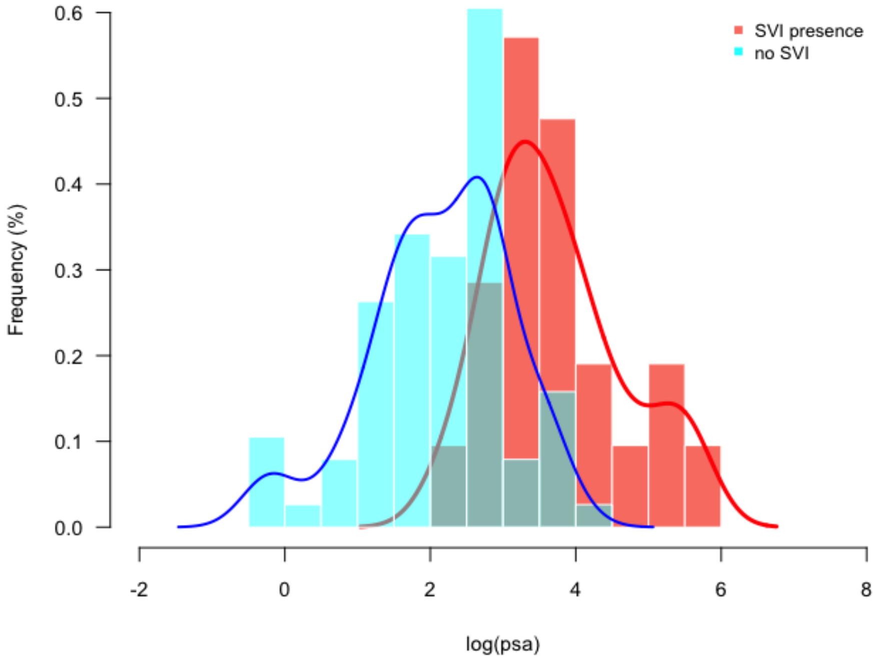
Traditionally, prostate cancer grades were described according to the Gleason Score. It ranges from 6 to 10. The lower the Gleason score, the more the cancer cells look like normal cells and are likely to grow and spread slowly. The Gleason score is used to help plan treatment.

**How many patients' Gleason score is 6?**

**5. Make a side-by-side boxplot of log(psa) for svi groups. Customize it to have an advanced graph (visually appealing, easy to read and understand).**

**6. Make an overlapping histogram (two histograms in the same graph) of log(psa) for svi groups. Customize it to have an advanced graph.** You may want to Google search for *overlapping histograms in R*.

Make your graph similar to the following.



**7. Summarize the lpsa values of the patients for both svi groups.**

**8. Describe the distribution of lpsa values of the patients for both svi groups using questions 5-7 (that is, side-by-side box plot, histograms, and summary statistics).**