موسسه آموزش عالی آزاد توسعه

برگزار کننده دوره‌های تخصصی علم داده

Homework 1
Please email your HWs to y.zerehsaz@gmail.com

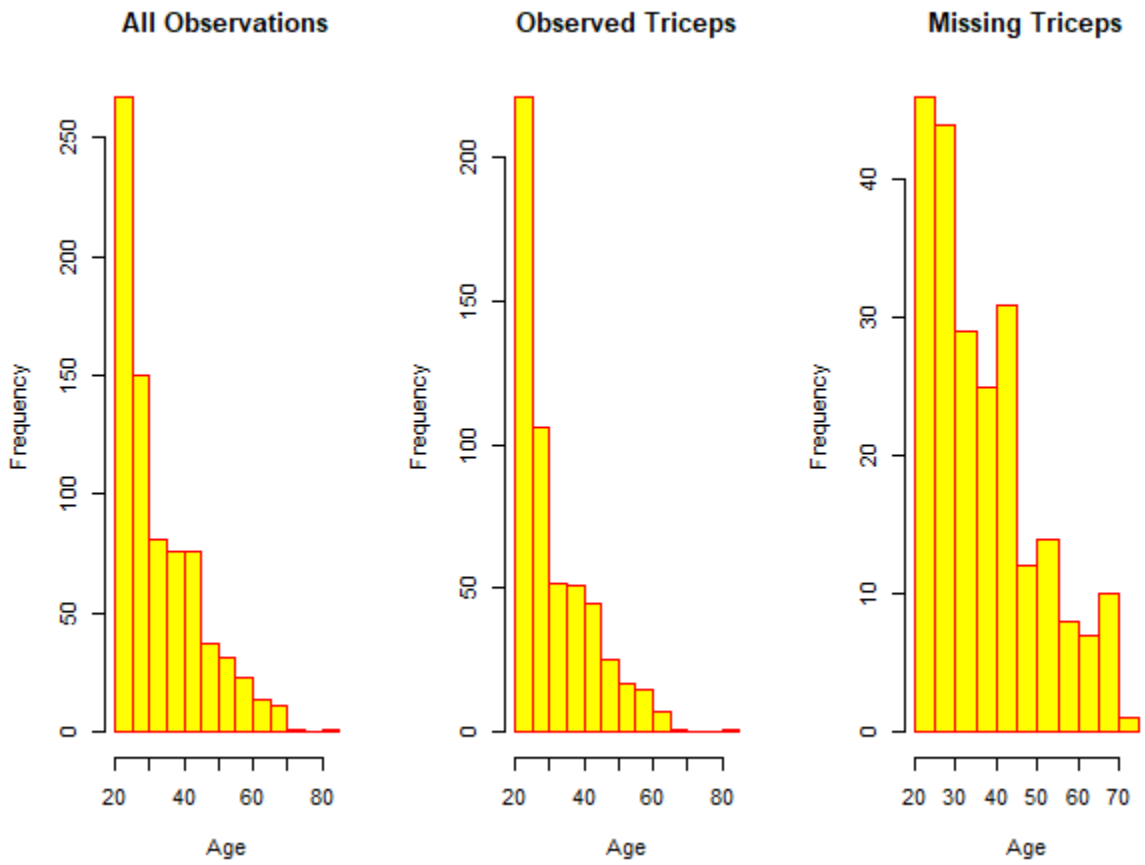
*****Please hand in your HWs as a word file with your email as the document's name. For instance, I would name my word file as y.zerehsaz@gmail.com.docx.**
*****Make sure to copy and paste the codes that you used for each question. I need to see your plots, results and conclusions but not the long output of your codes.**
*****When asked, please explain your results.**

Read data pima from the library "faraway", run the following codes and answer the questions.

```
library(faraway)
data(pima)
d<-pima
d$diastolic[d$diastolic==0]=NA
d$glucose[d$glucose==0]=NA
d$triceps[d$triceps==0]=NA
d$bmi[d$bmi==0]=NA
d$insulin[d$insulin==0]=NA
```

1- Plot the histogram of age for both observed and missing triceps. The functions that you might need are hist(), is.na(), is.finite(), d$triceps and d$age. Please explain the results.

```
par(mfrow=c(1,3))
hist(d$age,col="yellow",border="red",main="All Observations",xlab="Age")
hist(d$age[is.finite(d$triceps)],col="yellow",border="red",main="Observed Tri
ceps",xlab="Age")
hist(d$age[is.na(d$triceps)],col="yellow",border="red",main="Missing Triceps"
,xlab="Age")
```



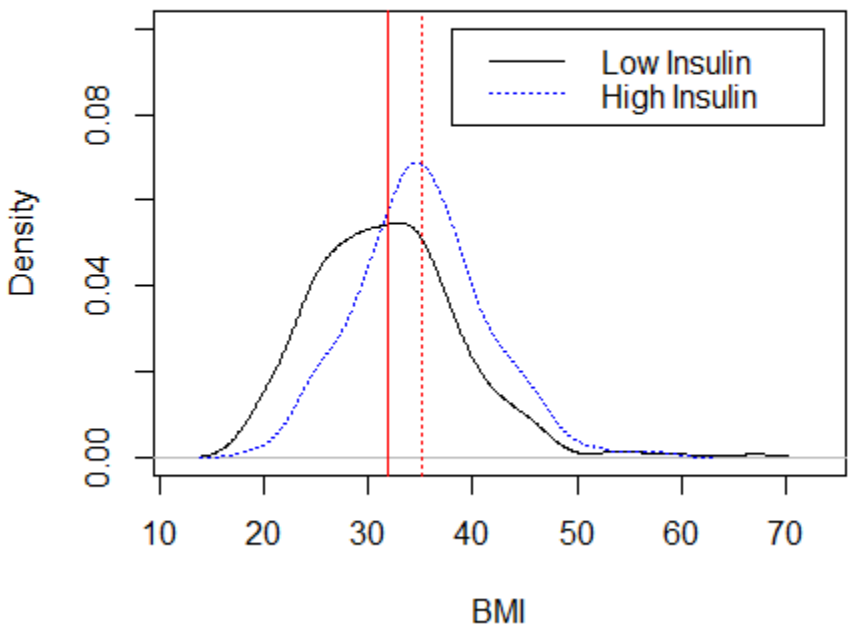
We should always be cautious when interpreting the results specially for datasets out of the scope of our expertise. But, we can make some conjectures, and then verify our hypotheses. As it can be seen, the age variable behaves more evenly when triceps values are missing. When triceps are observed, the behavior of age is more similar to that of age for all values of triceps. Furthermore, the average of age seems to be higher for missing triceps. One potential reason can be the unreliability of triceps as measures of body fat.

2- Compute the average of BMI for people with insulin levels lower and more than the average (I mean insulin average). So, you must calculate two averages for BMI. To get the average of inslulin, for instance, you can run this code

mean(d$insulin,na.rm=T)

Do not forget to set the argument na.rm to T.

```
mean(d$bmi[d$insulin>mean(d$insulin,na.rm=T)],na.rm=T)
[1] 35.07415
> mean(d$bmi[d$insulin<mean(d$insulin,na.rm=T)],na.rm=T)
[1] 31.87642
```



The average of BMI seems to be larger for people with above-the-average insulin. This can be justified based on the fact that insulin resistance is higher for people with large BMI. This means that body does not respond properly to the insulin which it makes. As a result, the blood will have large values of sugar.

Chung JO, Cho DH, Chung DJ, Chung MY. Associations among body mass index, insulin resistance, and pancreatic β-cell function in Korean patients with new-onset type 2 diabetes. *Korean J Intern Med*. 2012;27(1):66–71. doi:10.3904/kjim.2012.27.1.66

3- We need to compute confidence intervals for insulin variable for both levels of "test" variable. Remember that d$test gives you the diabetes test results for all subjects with 1 as a positive result and 0 representing a negative result. To do this, first, we are going to need to remove the missing values (only for simplicity) from the insulin variable. So, we can define a new variable called Insulin as

Insulin = d$insulin[is.finite(d$insulin)]

Now, use the "tapply" function (or any other function which you like) to compute the means and standard deviations of Insulin (Please see Slide 17 or type ?tapply in R console for more information). Just remember that d$test needs to be modified according to the new variable Insulin. So, we need to keep those elements in d$test associated with only the *observed* elements in d$insulin.

After computing the averages and standard deviations for Insulin with regard to both levels of d$test, we can now build 95% CIs for means of Insulin as

$$\bar{I}_i \pm 1.96\frac{S_i}{\sqrt{n_i}}; i = 1,2$$

where $\bar{I}_i$ is the average of Insulin computed for the ith level of d$test, S_i is the standard deviation of Insulin for level i, and n_i gives the sample size in group i.

Do the CIs overlap? What does this mean?

Potentially helpful functions include tapply, mean, sd (standard deviation), sqrt (square root) and length.

```
cbind(tapply(Insulin,Test,mean)-1.96*tapply(Insulin,Test,sd)/sqrt(tapply(Insu
lin,Test,length)),tapply(Insulin,Test,mean)+1.96*tapply(Insulin,Test,sd)/sqrt
(tapply(Insulin,Test,length)))
      [,1]     [,2]
0 117.9255 142.6503
1 184.0346 229.6577
```

The confidence intervals computed for mean of insulin do not overlap for these two groups. This means that the mean of insulin for people with diabetes is significantly larger than those without diabetes. Note that this ***does not mean*** the mean of insulin is between 117 and 142 with 95% probability for people without diabetes. This means if we repeat the same procedure several times, the mean of insulin for each group will fall in its associated interval 95% of times. To justify this phenomenon, we can refer to the fact that for Type-2 diabetes, insulin malfunctions in absorbing and transferring glucose from blood to muscles. This causes the pancreas to produce more insulin; thus increasing insulin levels in body.