EE_Graded_Lab_2

```
#graded lab 3: Evan Edelstein

#Load packages
library(MASS)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(doBy)
```

```
##
## Attaching package: 'doBy'
```

```
## The following object is masked from 'package:dplyr':
##
##     order_by
```

```
# 1) load data set
df_source <- read.csv("/Users/user/Desktop/School/2021/Spring2021/BTM-6000/Module9/Graded Lab 3.csv")
# 2a) inspect dataset
ls(df_source)
```

```
## [1] "age"        "CKD"        "DBP"        "eGFR"       "gender"
## [6] "Patient_ID" "SBP"
```

```
summary(df_source$CKD)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.000   0.000   0.844   2.000   5.000      77
```

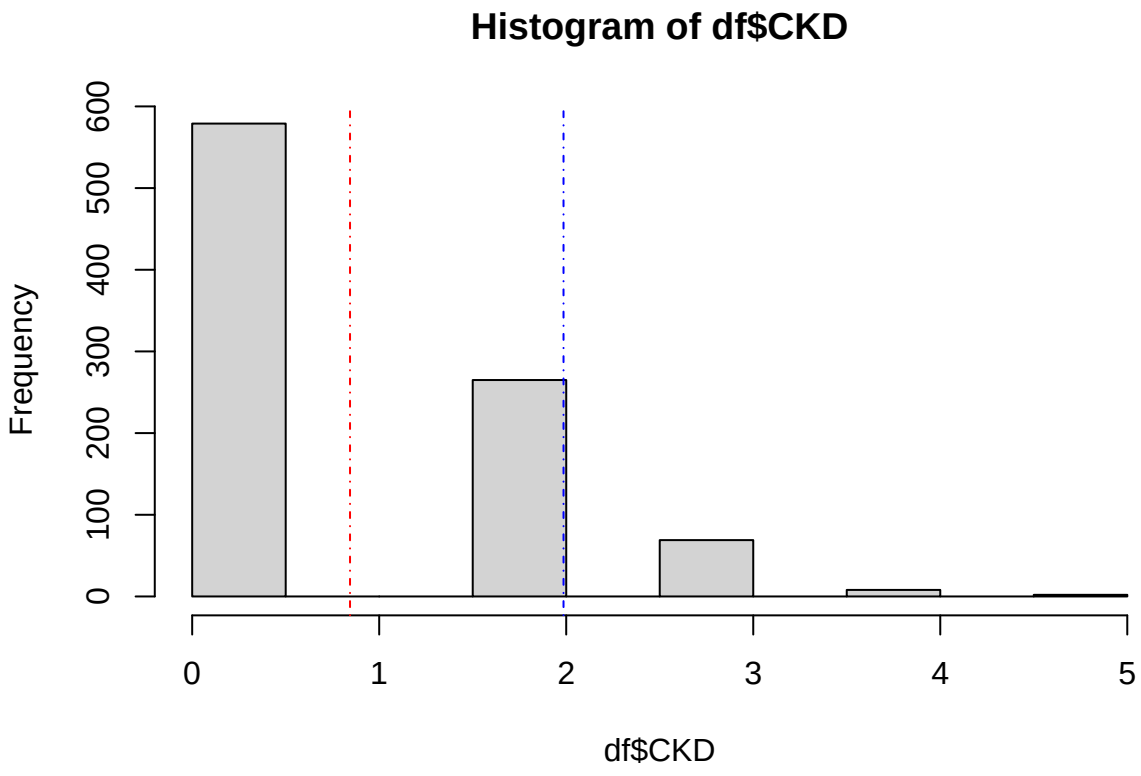
```
# 2b) remove missing values
df<-df_source[which(!is.na(df_source$CKD)), ]
# 2c) how many missing did I exclude
missing <- nrow(df_source) - nrow(df)
missing
```

```
## [1] 77
```

```
# 3)describe data set and build hist with mean(red) and sd(blue)
summary(df)
```

```
##    Patient_ID          age             gender           SBP
##  Min.   :   29   Min.   :18.00   Min.   :1.000   Min.   : 84.0
##  1st Qu.:19443   1st Qu.:30.00   1st Qu.:1.000   1st Qu.:110.0
##  Median :35805   Median :45.00   Median :2.000   Median :122.0
##  Mean   :33953   Mean   :46.52   Mean   :1.525   Mean   :124.7
##  3rd Qu.:49364   3rd Qu.:62.50   3rd Qu.:2.000   3rd Qu.:134.0
##  Max.   :62003   Max.   :85.00   Max.   :2.000   Max.   :230.0
##       DBP              CKD             eGFR
##  Min.   :  0.00   Min.   :0.000   Min.   :  5.40
##  1st Qu.: 62.00   1st Qu.:0.000   1st Qu.: 79.99
##  Median : 70.00   Median :0.000   Median : 98.92
##  Mean   : 69.73   Mean   :0.844   Mean   : 97.26
##  3rd Qu.: 78.00   3rd Qu.:2.000   3rd Qu.:116.75
##  Max.   :128.00   Max.   :5.000   Max.   :165.44
```

```
hist(df$CKD)
abline(v=mean(df$CKD),col="red",cex=1.2,lty=4)
abline(v=mean(df$CKD)+sd(df$CKD),col="blue",cex=1.2,lty=4)
abline(v=mean(df$CKD)-sd(df$CKD),col="blue",cex=1.2,lty=4)
```




```
# 4a) split dataset by gender
df$MALE<-ifelse(df$gender==1,1,0)
sum(df$MALE)
```

```
## [1] 438
```

```
df$MALE<-factor(df$MALE,levels=c(0,1),labels=c("female","male"))
summary(df$MALE)
```

```
## female   male
##    485    438
```

```
# 4b) split by ckd stage greater or equal to 3, those who have ckd stage 3 or higher are considered to
df$CKD3<-ifelse(df$CKD>=3,1,0)
sum(df$CKD3)
```

```
## [1] 79
```

```
df$CKD3<-factor(df$CKD3,levels=c(0,1),labels=c("no_CKD","CKD"))
summary(df$CKD3)
```

```
## no_CKD    CKD
##    844     79
```

```
# 4c) make output frame and populate
demographics<-data.frame("Parameter"=c("sample_size","age","SBP","DBP","eGFR") ,"CKD_all"=NA,"no_CKD_al
#  add sample size to first row
demographics[ 1, 2]<-length(df$Patient_ID[df$CKD3=="CKD"])
demographics[ 1, 3]<-length(df$Patient_ID[df$CKD3=="no_CKD"])
demographics[ 1, 4]<-length(df$Patient_ID[df$CKD3=="no_CKD" & df$MALE=="male"])
demographics[ 1, 5]<-length(df$Patient_ID[df$CKD3=="no_CKD" & df$MALE=="female"])
demographics[ 1, 6]<-length(df$Patient_ID[df$CKD3=="CKD" & df$MALE=="male"])
demographics[ 1, 7]<-length(df$Patient_ID[df$CKD3=="CKD" & df$MALE=="female"])


# 4) populate ages
demographics[ 2, 2]<-paste(round(mean(df$age[df$CKD3 =="CKD"]),1),"+/-",round(sd(df$age[df$CKD3 =="CKD"
demographics[ 2, 3]<-paste(round(mean(df$age[df$CKD3 =="no_CKD"]),1),"+/-",round(sd(df$age[df$CKD3 =="n
demographics[ 2, 4]<-paste(round(mean(df$age[df$CKD3 =="no_CKD" & df$MALE == "male"]),1),"+/-",round(sd
demographics[ 2, 5]<-paste(round(mean(df$age[df$CKD3 =="no_CKD" & df$MALE == "female"]),1),"+/-",round(
demographics[ 2, 6]<-paste(round(mean(df$age[df$CKD3 =="CKD" & df$MALE == "male"]),1),"+/-",round(sd(df
demographics[ 2, 7]<-paste(round(mean(df$age[df$CKD3 =="CKD" & df$MALE == "female"]),1),"+/-",round(sd(

# 4) populate sbp
demographics[ 3, 2]<-paste(round(mean(df$SBP[df$CKD3 =="CKD"]),1),"+/-",round(sd(df$SBP[df$CKD3 =="CKD"
demographics[ 3, 3]<-paste(round(mean(df$SBP[df$CKD3 =="no_CKD"]),1),"+/-",round(sd(df$SBP[df$CKD3 =="n
demographics[ 3, 4]<-paste(round(mean(df$SBP[df$CKD3 =="no_CKD" & df$MALE == "male"]),1),"+/-",round(sd
demographics[ 3, 5]<-paste(round(mean(df$SBP[df$CKD3 =="no_CKD" & df$MALE == "female"]),1),"+/-",round(
demographics[ 3, 6]<-paste(round(mean(df$SBP[df$CKD3 =="CKD" & df$MALE == "male"]),1),"+/-",round(sd(df
demographics[ 3, 7]<-paste(round(mean(df$SBP[df$CKD3 =="CKD" & df$MALE == "female"]),1),"+/-",round(sd(

# 4)populate DBP
demographics[ 4, 2]<-paste(round(mean(df$DBP[df$CKD3 =="CKD"]),1),"+/-",round(sd(df$DBP[df$CKD3 =="CKD"
demographics[ 4, 3]<-paste(round(mean(df$DBP[df$CKD3 =="no_CKD"]),1),"+/-",round(sd(df$DBP[df$CKD3 =="n
demographics[ 4, 4]<-paste(round(mean(df$DBP[df$CKD3 =="no_CKD" & df$MALE == "male"]),1),"+/-",round(sd
demographics[ 4, 5]<-paste(round(mean(df$DBP[df$CKD3 =="no_CKD" & df$MALE == "female"]),1),"+/-",round(
demographics[ 4, 6]<-paste(round(mean(df$DBP[df$CKD3 =="CKD" & df$MALE == "male"]),1),"+/-",round(sd(df
demographics[ 4, 7]<-paste(round(mean(df$DBP[df$CKD3 =="CKD" & df$MALE == "female"]),1),"+/-",round(sd(

# 4)populate eGFR
demographics[ 5, 2]<-paste(round(mean(df$eGFR[df$CKD3 =="CKD"]),1),"+/-",round(sd(df$eGFR[df$CKD3 =="CK
demographics[ 5, 3]<-paste(round(mean(df$eGFR[df$CKD3 =="no_CKD"]),1),"+/-",round(sd(df$eGFR[df$CKD3 ==
demographics[ 5, 4]<-paste(round(mean(df$eGFR[df$CKD3 =="no_CKD" & df$MALE == "male"]),1),"+/-",round(sd
demographics[ 5, 5]<-paste(round(mean(df$eGFR[df$CKD3 =="no_CKD" & df$MALE == "female"]),1),"+/-",round
demographics[ 5, 6]<-paste(round(mean(df$eGFR[df$CKD3 =="CKD" & df$MALE == "male"]),1),"+/-",round(sd(d
```

```r
demographics[ 5, 7]<-paste(round(mean(df$eGFR[df$CKD3 =="CKD" & df$MALE == "female"]),1),"+/-",round(sd



# 5a) Generate the research questions
# a) null -> there is no difference in CKD for men and woman
# b) alt. -> there is a statistically signifigent difference in CKD patiants between male and femlales.

# 5b)
#a) null -> there is no difference between the SBP between CKD pataints and non-CKD pataints in the mal
#b) alt. -> there is a signifigant difference between the SBP between CKD pataints and non-CKD pataints


# 6 & 7)populate diff column and pval for continous variables for 1. all 2. men and 3. woman

# add p-value column
demographics$P_all_CKD_noCKD <- NA

#Loop through "all" data using iterator "row"
for (row in 2:nrow(demographics)) {
  var <- demographics[row,1] #<- var is the variable to work on, [age,SBP, DBP,eGFR]
  ttest<-t.test(df[[var]][df$CKD3 =="no_CKD"],df[[var]][df$CKD3 =="CKD"]) #<- perform t-test
  diff_age_all<-ttest$estimate[1]-ttest$estimate[2]  #<- in difference in ttest
  LL_CI_age_all<-ttest$conf.int[1] # <- lower level for confidence interval
  UL_CI_age_all<-ttest$conf.int[2] #<- upper level for CI
  pval <- ttest$p.value #<- p-value
  # add values to demographics
  demographics[row,8]<-paste(round(diff_age_all,1)," (",round(LL_CI_age_all,1)," to ",round(UL_CI_age_al
  demographics[row,9] <- pval
}
#add column for male diff and male pvalue
demographics$Diff_men_CKD_minus_noCKD<- NA
demographics$P_men_CKD_noCKD <- NA
#Loop through "male" data using iterator "row"
for (row in 2:nrow(demographics)) {
  var <- demographics[row,1] #<- var is the variable to work on, [age,SBP, DBP,eGFR]
  # perform t-test <-> see above for comments
  ttest<-t.test(df[[var]][df$CKD3 =="no_CKD" & df$MALE=="male"],df[[var]][df$CKD3 =="CKD" & df$MALE=="ma
  diff_var<-ttest$estimate[1]-ttest$estimate[2]
  LL_CI_var<-ttest$conf.int[1]
  UL_CI_age_var<-ttest$conf.int[2]
  pval <- ttest$p.value
  #write to demographics
  demographics[row,10]<-paste(round(diff_var,1)," (",round(LL_CI_var,1)," to ",round(UL_CI_age_var,1),"
  demographics[row,11] <- pval

}
# create diff and pval for female
demographics$Diff_women_minus_CKD_noCKD <- NA
demographics$P_women_CKD_noCKD <-NA

#Loop through "female" data using iterator "row"
for (row in 2:nrow(demographics)) {
```

```
  var <- demographics[row,1]
  #perfrom t-test
  ttest<-t.test(df[[var]][df$CKD3 =="no_CKD"& df$MALE=="female"],df[[var]][df$CKD3 =="CKD"& df$MALE=="f
  diff_var<-ttest$estimate[1]-ttest$estimate[2]
  LL_CI_var<-ttest$conf.int[1]
  UL_CI_var<-ttest$conf.int[2]
  pval <- ttest$p.value
  #write to demographics
  demographics[row,12]<-paste(round(diff_var,1)," (",round(LL_CI_var,1)," to ",round(UL_CI_var,1),")",s
  demographics[row,13] <- pval

}

# 6 & 7) fill in  diff columns for non-cont. variables
demographics[ 1, 8] <- strtoi(demographics[ 1, 3]) - strtoi(demographics[ 1, 2])
demographics[ 1, 10] <- strtoi(demographics[ 1, 4]) - strtoi(demographics[ 1, 6])
demographics[ 1, 12] <- strtoi(demographics[ 1, 5]) - strtoi(demographics[ 1, 7])

#show demographics table and save
demographics
```

```
##     Parameter      CKD_all   no_CKD_all   no_CKD_men no_CKD_women      CKD_men
## 1 sample_size           79          844          404          440           34
## 2         age  45.1+/-17.2  46.6+/-19.4  46.5+/-18.8  46.8+/-20.1  45.8+/-19.5
## 3         SBP 123.1+/-16.4 124.8+/-20.4 127.1+/-18.2 122.8+/-22.1 126.1+/-17.3
## 4         DBP  70.8+/-10.8  69.6+/-15.1  72.9+/-14.9  66.6+/-14.6  74.1+/-11.9
## 5        eGFR    45.1+/-12 102.1+/-21.5 102.8+/-21.7 101.5+/-21.3    44.5+/-14
##      CKD_women           Diff_all P_all_CKD_noCKD Diff_men_CKD_minus_noCKD
## 1           45                765              NA                       370
## 2  44.6+/-15.4  1.5 (-2.6 to 5.6)    4.630312e-01        0.7 (-6.3 to 7.7)
## 3 120.8+/-15.5  1.8 (-2.1 to 5.7)    3.703889e-01        1.1 (-5.2 to 7.3)
## 4   68.3+/-9.3 -1.1 (-3.7 to 1.5)    3.948191e-01       -1.1 (-5.5 to 3.2)
## 5  45.5+/-10.4    57 (54 to 60.1)    3.119107e-71      58.3 (53.1 to 63.6)
##   P_men_CKD_noCKD Diff_women_minus_CKD_noCKD P_women_CKD_noCKD
## 1              NA                        395               NA
## 2    8.443188e-01          2.2 (-2.8 to 7.1)      3.913050e-01
## 3    7.370291e-01              2 (-3.1 to 7)      4.433539e-01
## 4    6.029177e-01         -1.7 (-4.7 to 1.4)      2.880520e-01
## 5    9.388451e-27          56 (52.3 to 59.6)      2.974108e-48
```

```
write.csv(demographics,"/Users/user/Desktop/School/2021/Spring2021/BTM-6000/Module9/EE_graded_lab9.csv")

# 8) contingency table
tbl<-table(df$CKD3,df$MALE)
tbl
```

```
##
##          female male
##   no_CKD    440  404
##   CKD        45   34
```

```
# 9) chi-sqr. test
chisq.test(tbl)
```

```
##
```

```
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl
## X-squared = 0.49586, df = 1, p-value = 0.4813
```

```
# pvalue greater than 0.05 so we fail to reject the null so the data suggests there is  a differance bt

demographics$Diff_in_P_val <- NA
for(row in 2:nrow(demographics)){
  val <- demographics[row,11] -  demographics[row,13]
  demographics[row,14]<- val

}

#the differance in pvals were all greater than 0.05 thus we fail to reject the null hypt.
#-> there is  a differance btw men and woman in terms of CKD
```