Project

Tharun Prabhakar

2023-11-10

library(NHANES)
library(dplyr)

```
library(corrplot)
library(doBy)
library(table1)
library(knitr)
library(papeR)
library(stats)
library(tidyverse)
library(ggplot2)
library(corrplot)
SD_pop \leftarrow function(x,n) \{ sqrt(sum((x - mean(x))^2)/(n)) \}
data(NHANES)
ls(NHANES)
    [1] "Age"
                             "Age1stBaby"
                                                 "AgeDecade"
                                                                     "AgeFirstMarij"
##
    [5] "AgeMonths"
                             "AgeRegMarij"
                                                 "Alcohol12PlusYr"
                                                                      "AlcoholDay"
##
                             "BMI"
  [9] "AlcoholYear"
                                                 "BMI_WHO"
                                                                      "BMICatUnder20yrs"
##
## [13] "BPDia1"
                             "BPDia2"
                                                 "BPDia3"
                                                                     "BPDiaAve"
## [17] "BPSys1"
                             "BPSys2"
                                                 "BPSys3"
                                                                     "BPSysAve"
## [21] "CompHrsDay"
                             "CompHrsDayChild"
                                                 "DaysMentHlthBad"
                                                                     "DaysPhysHlthBad"
## [25] "Depressed"
                             "Diabetes"
                                                                      "DirectChol"
                                                 "DiabetesAge"
## [29] "Education"
                             "Gender"
                                                 "HardDrugs"
                                                                      "HeadCirc"
## [33] "HealthGen"
                             "Height"
                                                 "HHIncome"
                                                                     "HHIncomeMid"
## [37] "HomeOwn"
                             "HomeRooms"
                                                 "TD"
                                                                      "Length"
## [41] "LittleInterest"
                             "Marijuana"
                                                 "MaritalStatus"
                                                                      "nBabies"
## [45] "nPregnancies"
                             "PhysActive"
                                                 "PhysActiveDays"
                                                                     "Poverty"
## [49] "PregnantNow"
                             "Pulse"
                                                 "Race1"
                                                                     "Race3"
                             "SameSex"
## [53] "RegularMarij"
                                                 "SexAge"
                                                                      "SexEver"
## [57] "SexNumPartnLife"
                             "SexNumPartYear"
                                                 "SexOrientation"
                                                                     "SleepHrsNight"
## [61] "SleepTrouble"
                             "Smoke100"
                                                 "Smoke100n"
                                                                     "SmokeAge"
## [65] "SmokeNow"
                             "SurveyYr"
                                                 "Testosterone"
                                                                      "TotChol"
## [69] "TVHrsDay"
                             "TVHrsDayChild"
                                                 "UrineFlow1"
                                                                      "UrineFlow2"
## [73] "UrineVol1"
                             "UrineVol2"
                                                 "Weight"
                                                                     "Work"
NHANES<-NHANES[!duplicated(NHANES$ID),]</pre>
df<-select(NHANES,ID,Gender,Age,BMI,HHIncomeMid,UrineFlow1)</pre>
df<-df[complete.cases(df),]</pre>
df<-df %>% filter(Age > 18)
```

```
if (df$BMI[i]<18.5) {</pre>
    df$BMI_Factor[i]<-"Under Weight"</pre>
  else if(df$BMI[i]>=18.5 & df$BMI[i]<25){</pre>
    df$BMI_Factor[i]<-"Normal"</pre>
  else if(df$BMI[i]>=25 & df$BMI[i]<30){</pre>
    df$BMI_Factor[i]<-"Over Weight"</pre>
  else if(df$BMI[i]>=30){
    df$BMI_Factor[i]<-"Obese"</pre>
}
## Warning: Unknown or uninitialised column: 'BMI Factor'.
df$BMI_Factor<-factor(df$BMI_Factor,order=TRUE,</pre>
                       levels = c("Under Weight", "Normal", "Over Weight", "Obese"))
df$Income_Factor <- cut(df$HHIncomeMid,breaks = c(min(df$HHIncomeMid)-1,22500,87500,max(df$HHIncomeMid)
df$Income_Factor <-factor(df$Income_Factor,order=TRUE,levels=c("Low","Medium","High"))</pre>
summary(df$Income_Factor)
##
      Low Medium
                    High
##
     1039
            2109
                     845
set.seed(25)
selection<-sample(df$ID[df$Age > 18],500)
df1<-data.frame("ID"=df$ID[df$ID %in% selection],
                 "age"=df$Age[df$ID %in% selection],
                 "Gender"=df$Gender[df$ID %in% selection],
                 "BMI"=df$BMI[df$ID %in% selection],
                 "BMI_Factor"=df$BMI_Factor[df$ID %in% selection],
                 "Urine_Flow"=df$UrineFlow1[df$ID %in% selection],
                 "Income"=df$HHIncomeMid[df$ID %in% selection],
                 "Income_Factor"=df$Income_Factor[df$ID %in% selection])
```

The Summary Statistics of the Population:

kable(summarize(df))

for (i in 1:length(df\$BMI))

	N	Mean	SD	Min	Q1	Median	Q3	Max
ID	3993	61678.99	5874.41	51630.00	56664.00	61575.00	66822.00	71915.00
Age	3993	46.79	17.53	19.00	32.00	45.00	60.00	80.00
BMI	3993	28.82	6.72	15.02	24.10	27.70	32.20	81.25
HHIncomeN	Ai & 993	54256.82	32903.22	2500.00	22500.00	50000.00	87500.00	100000.00
Urine Flow 1	3993	1.01	0.94	0.00	0.43	0.74	1.26	12.35

kable(summarize(df, type = "factor",cumulative=TRUE))

	Level	N	%	∑ %
Gender	female	2041	51.1	51.1
	male	1952	48.9	100.0
BMI_Factor	Under Weight	69	1.7	1.7
	Normal	1146	28.7	30.4
	Over Weight	1341	33.6	64.0
	Obese	1437	36.0	100.0
$Income_Factor$	Low	1039	26.0	26.0
	Medium	2109	52.8	78.8
	High	845	21.2	100.0

The Summary Statistics of the Sample:

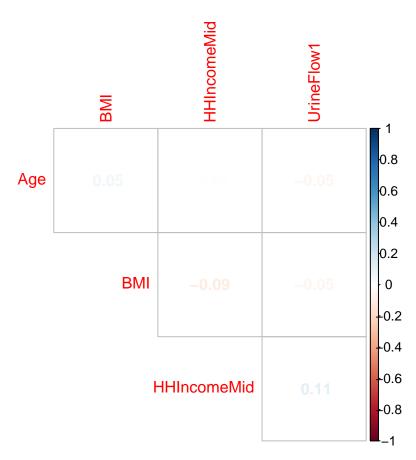
kable(summarize(df1))

-	N	Mean	SD	Min	Q1	Median	Q3	Max
ID	500	61760.50	5825.41	51654.00	56973.50	61526.50	66925.50	71875.00
age	500	46.98	17.36	19.00	33.00	46.00	61.00	80.00
$\overline{\mathrm{BMI}}$	500	29.20	7.12	15.97	24.27	28.44	32.80	81.25
$Urine_Flo$	w 500	1.04	1.10	0.00	0.45	0.72	1.23	12.35
Income	500	57515.00	33449.59	2500.00	30000.00	50000.00	100000.00	100000.00

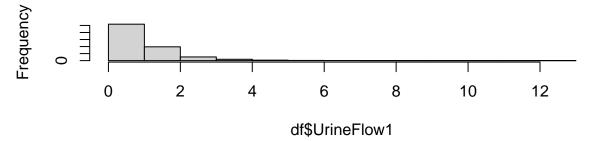
kable(summarize(df1, type = "factor",cumulative=TRUE))

	Level	N	%	Σ %
Gender	female	253	50.6	50.6
	male	247	49.4	100.0
BMI_Factor	Under Weight	6	1.2	1.2
	Normal	146	29.2	30.4
	Over Weight	151	30.2	60.6
	Obese	197	39.4	100.0
$Income_Factor$	Low	116	23.2	23.2
	Medium	257	51.4	74.6
	High	127	25.4	100.0

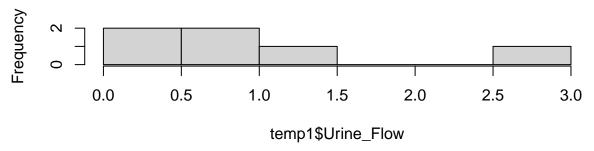
[1] -0.04621187



```
# z-test
temp1<-df1[df1$BMI_Factor=='Under Weight',]</pre>
summary(temp1$Urine_Flow)
##
      Min. 1st Qu. Median
                              Mean 3rd Qu.
                                              Max.
  0.0910 0.4150 0.7205 0.9295 0.9600 2.6920
summary(df$UrineFlow1)
##
     Min. 1st Qu. Median
                              Mean 3rd Qu.
                                              Max.
##
     0.000
            0.430
                     0.739
                             1.006
                                     1.261 12.346
par(mfrow=c(2,1))
hist(df$UrineFlow1)
hist(temp1$Urine_Flow)
```



Histogram of temp1\$Urine_Flow



[1] -0.2002546

```
#2-tailed
p_2tailed<-2 * pnorm( - abs(z))
p_2tailed</pre>
```

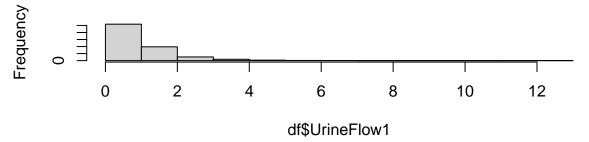
```
#not rejecting the null

#1-tailed

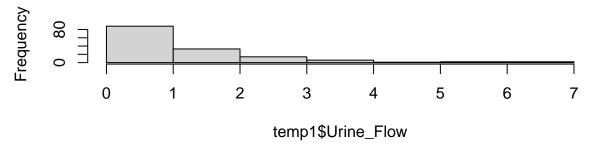
#Null hypothesis sample is greater/equal than the mew

#Research hypothesis sample is smaller than the mew
```

```
p_1tailed_smaller <- pnorm(z)</pre>
p_1tailed_smaller
## [1] 0.4206407
#not rejecting the null
#1-tailed
#Null hypothesis sample is smaller/equal than the mew
\#Research\ hypothesis\ sample\ is\ greater\ than\ the\ mew
p_1tailed_greater <- 1 - pnorm(z)</pre>
p_1tailed_greater
## [1] 0.5793593
#not rejecting the null
# z-test
temp1<-df1[df1$BMI_Factor=='Normal',]</pre>
summary(temp1$Urine_Flow)
##
      Min. 1st Qu. Median
                              Mean 3rd Qu.
                                               Max.
## 0.0970 0.4248 0.7980 1.1925 1.5835 6.9000
summary(df$UrineFlow1)
##
      Min. 1st Qu. Median
                              Mean 3rd Qu.
                                               Max.
##
     0.000 0.430 0.739
                             1.006 1.261 12.346
par(mfrow=c(2,1))
hist(df$UrineFlow1)
hist(temp1$Urine_Flow)
```



Histogram of temp1\$Urine_Flow



[1] 2.390759

```
#2-tailed
p_2tailed<-2 * pnorm( - abs(z))
p_2tailed</pre>
```

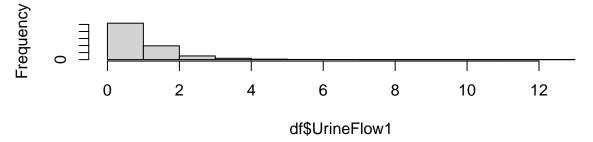
```
#not rejecting the null

#1-tailed

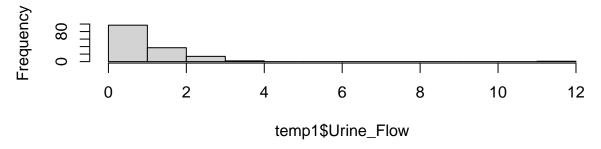
#Null hypothesis sample is greater/equal than the mew

#Research hypothesis sample is smaller than the mew
```

```
p_1tailed_smaller <- pnorm(z)</pre>
p_1tailed_smaller
## [1] 0.9915932
#not rejecting the null
#1-tailed
#Null hypothesis sample is smaller/equal than the mew
\#Research\ hypothesis\ sample\ is\ greater\ than\ the\ mew
p_1tailed_greater <- 1 - pnorm(z)</pre>
p_1tailed_greater
## [1] 0.008406796
#not rejecting the null
# z-test
temp1<-df1[df1$BMI_Factor=='Over Weight',]</pre>
summary(temp1$Urine_Flow)
##
     Min. 1st Qu. Median Mean 3rd Qu.
                                              Max.
     0.051 0.464 0.731 1.007 1.224 11.069
##
summary(df$UrineFlow1)
##
      Min. 1st Qu. Median
                              Mean 3rd Qu.
                                              Max.
##
     0.000 0.430 0.739
                             1.006 1.261 12.346
par(mfrow=c(2,1))
hist(df$UrineFlow1)
hist(temp1$Urine_Flow)
```



Histogram of temp1\$Urine_Flow



[1] 0.001591493

```
#2-tailed
p_2tailed<-2 * pnorm( - abs(z))
p_2tailed</pre>
```

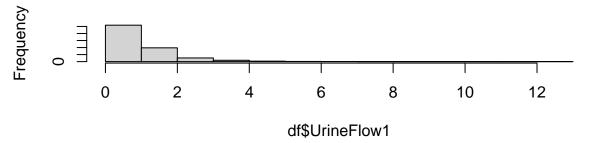
```
#not rejecting the null

#1-tailed

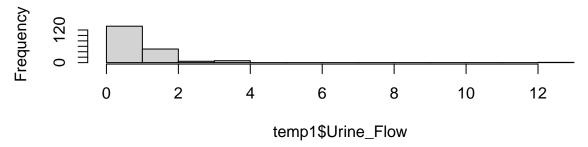
#Null hypothesis sample is greater/equal than the mew

#Research hypothesis sample is smaller than the mew
```

```
p_1tailed_smaller <- pnorm(z)</pre>
p_1tailed_smaller
## [1] 0.5006349
#not rejecting the null
#1-tailed
#Null hypothesis sample is smaller/equal than the mew
\#Research\ hypothesis\ sample\ is\ greater\ than\ the\ mew
p_1tailed_greater <- 1 - pnorm(z)</pre>
p_1tailed_greater
## [1] 0.4993651
#not rejecting the null
# z-test
temp1<-df1[df1$BMI_Factor=='Obese',]</pre>
summary(temp1$Urine_Flow)
##
      Min. 1st Qu. Median
                              Mean 3rd Qu.
                                               Max.
## 0.0000 0.4600 0.6860 0.9588 1.1690 12.3460
summary(df$UrineFlow1)
##
      Min. 1st Qu. Median
                              Mean 3rd Qu.
                                               Max.
##
     0.000 0.430 0.739
                             1.006 1.261 12.346
par(mfrow=c(2,1))
hist(df$UrineFlow1)
hist(temp1$Urine_Flow)
```



Histogram of temp1\$Urine_Flow



[1] -0.710129

```
#2-tailed
p_2tailed<-2 * pnorm( - abs(z))
p_2tailed</pre>
```

```
#not rejecting the null

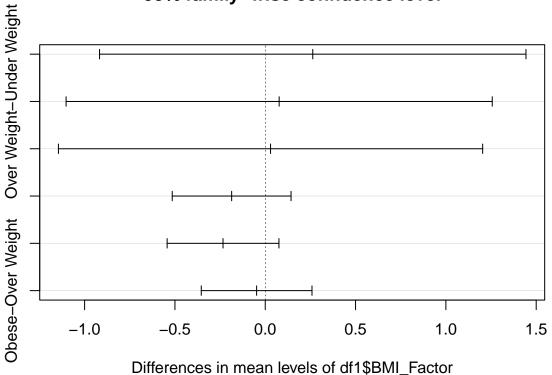
#1-tailed

#Null hypothesis sample is greater/equal than the mew

#Research hypothesis sample is smaller than the mew
```

```
p_1tailed_smaller <- pnorm(z)</pre>
p_1tailed_smaller
## [1] 0.2388121
#not rejecting the null
#1-tailed
#Null hypothesis sample is smaller/equal than the mew
{\it \#Research\ hypothesis\ sample\ is\ greater\ than\ the\ mew}
p_1tailed_greater <- 1 - pnorm(z)</pre>
p_1tailed_greater
## [1] 0.7611879
#not rejecting the null
anova_test<-aov(df1$Urine_Flow~df1$BMI_Factor)</pre>
summary(anova_test)
                    Df Sum Sq Mean Sq F value Pr(>F)
## df1$BMI_Factor
                                        1.362 0.254
                   3 4.9 1.646
## Residuals
                   496 599.4
                                1.208
Tukey<-TukeyHSD(anova_test)</pre>
plot(Tukey)
```



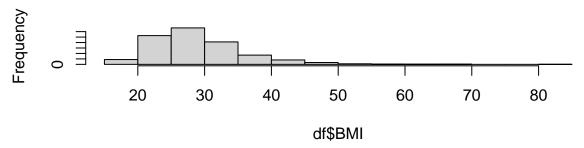


Tukey

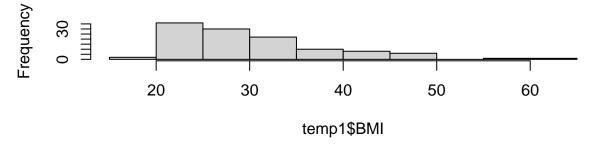
```
Tukey multiple comparisons of means
##
##
       95% family-wise confidence level
##
## Fit: aov(formula = df1$Urine_Flow ~ df1$BMI_Factor)
##
## $'df1$BMI_Factor'
##
                                    diff
                                                lwr
                                                            upr
                                                                    p adj
## Normal-Under Weight
                              0.26304110 -0.9173331 1.44341529 0.9396922
## Over Weight-Under Weight 0.07702980 -1.1025727 1.25663233 0.9983025
## Obese-Under Weight
                             0.02931218 -1.1450154 1.20363973 0.9999043
## Over Weight-Normal
                             -0.18601129 -0.5149109 0.14288826 0.4639303
## Obese-Normal
                             -0.23372891 -0.5431768 0.07571895 0.2098848
## Obese-Over Weight
                             -0.04771762 -0.3542088 0.25877358 0.9781094
# z-test
temp1<-df1[df1$Income_Factor=='Low',]</pre>
summary(temp1$BMI)
##
      Min. 1st Qu.
                    Median
                              Mean 3rd Qu.
                                               Max.
##
     16.70
             24.29
                     27.82
                              30.09
                                      33.64
                                              61.01
```

```
summary(df$BMI)
##
      Min. 1st Qu.
                    Median
                               Mean 3rd Qu.
                                               Max.
##
     15.02
             24.10
                     27.70
                              28.82
                                      32.20
                                              81.25
par(mfrow=c(2,1))
hist(df$BMI)
hist(temp1$BMI)
```

Histogram of df\$BMI



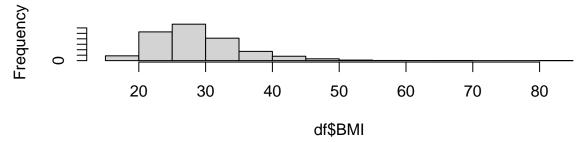
Histogram of temp1\$BMI



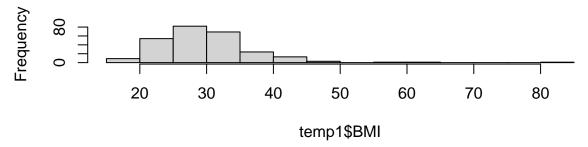
[1] 2.035792

```
#2-tailed
p_2tailed<-2 * pnorm( - abs(z))</pre>
p_2tailed
## [1] 0.0417712
#not rejecting the null
#1-tailed
#Null hypothesis sample is greater/equal than the mew
\#Research\ hypothesis\ sample\ is\ smaller\ than\ the\ mew
p_1tailed_smaller <- pnorm(z)</pre>
p_1tailed_smaller
## [1] 0.9791144
#not rejecting the null
#1-tailed
#Null hypothesis sample is smaller/equal than the mew
#Research hypothesis sample is greater than the mew
p_1tailed_greater <- 1 - pnorm(z)</pre>
p_1tailed_greater
## [1] 0.0208856
#rejecting the null
# z-test
temp1<-df1[df1$Income_Factor=='Medium',]</pre>
summary(temp1$BMI)
      Min. 1st Qu. Median Mean 3rd Qu.
##
                                               Max.
     17.60
           25.10 29.10
                             29.87 33.50
                                              81.25
summary(df$BMI)
##
      Min. 1st Qu. Median
                              Mean 3rd Qu.
                                               Max.
##
     15.02 24.10
                    27.70
                             28.82 32.20
                                              81.25
par(mfrow=c(2,1))
hist(df$BMI)
hist(temp1$BMI)
```

Histogram of df\$BMI



Histogram of temp1\$BMI



[1] 2.493383

```
#2-tailed
p_2tailed<-2 * pnorm( - abs(z))
p_2tailed</pre>
```

```
#not rejecting the null

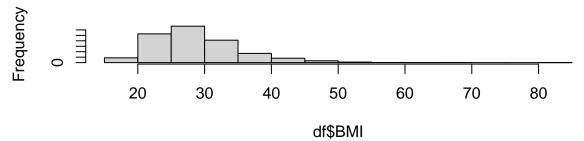
#1-tailed

#Null hypothesis sample is greater/equal than the mew

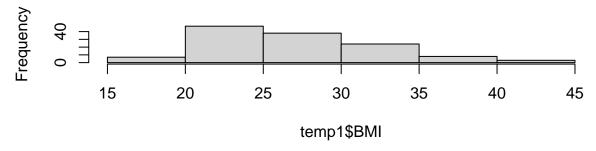
#Research hypothesis sample is smaller than the mew
```

```
p_1tailed_smaller <- pnorm(z)</pre>
p_1tailed_smaller
## [1] 0.9936734
#not rejecting the null
#1-tailed
#Null hypothesis sample is smaller/equal than the mew
\#Research\ hypothesis\ sample\ is\ greater\ than\ the\ mew
p_1tailed_greater <- 1 - pnorm(z)</pre>
p_1tailed_greater
## [1] 0.006326614
#rejecting the null
# z-test
temp1<-df1[df1$Income_Factor=='High',]</pre>
summary(temp1$BMI)
##
     Min. 1st Qu. Median Mean 3rd Qu.
                                              Max.
     15.97 23.09 26.70 27.04 30.23
                                             43.50
##
summary(df$BMI)
##
     Min. 1st Qu. Median
                             Mean 3rd Qu.
                                              Max.
##
     15.02 24.10
                    27.70
                             28.82
                                     32.20
                                             81.25
par(mfrow=c(2,1))
hist(df$BMI)
hist(temp1$BMI)
```

Histogram of df\$BMI



Histogram of temp1\$BMI



[1] -2.987004

```
#2-tailed
p_2tailed<-2 * pnorm( - abs(z))
p_2tailed</pre>
```

```
#not rejecting the null

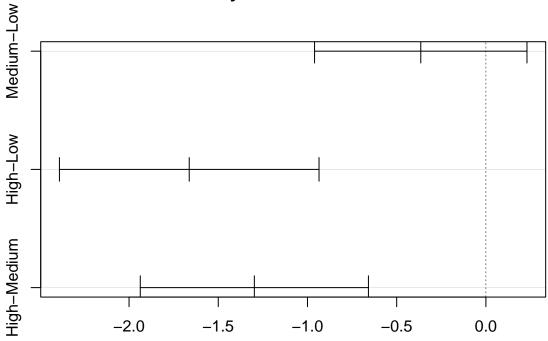
#1-tailed

#Null hypothesis sample is greater/equal than the mew

#Research hypothesis sample is smaller than the mew
```

```
p_1tailed_smaller <- pnorm(z)</pre>
p_1tailed_smaller
## [1] 0.001408628
#not rejecting the null
#1-tailed
#Null hypothesis sample is smaller/equal than the mew
\#Research\ hypothesis\ sample\ is\ greater\ than\ the\ mew
p_1tailed_greater <- 1 - pnorm(z)</pre>
p_1tailed_greater
## [1] 0.9985914
#rejecting the null
anova_test1<-aov(df$BMI~df$Income_Factor)</pre>
summary(anova_test1)
                      Df Sum Sq Mean Sq F value Pr(>F)
## df$Income_Factor 2 1433 716.3 15.96 1.24e-07 ***
## Residuals
                  3990 179019
                                   44.9
## ---
## Signif. codes: 0 '*** 0.001 '** 0.01 '* 0.05 '.' 0.1 ' ' 1
Tukey1<-TukeyHSD(anova_test1)</pre>
plot(Tukey1)
```

95% family-wise confidence level



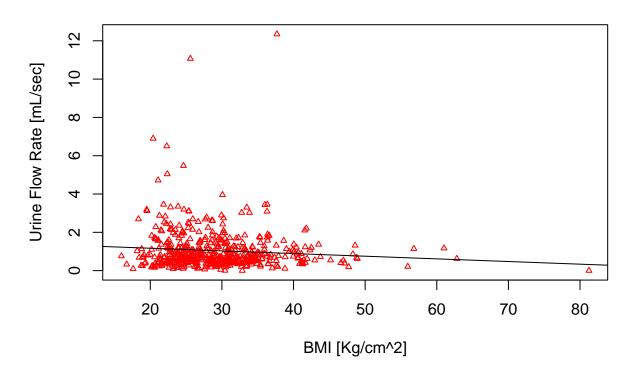
Differences in mean levels of df\$Income_Factor

Tukey1

```
Tukey multiple comparisons of means
##
##
       95% family-wise confidence level
## Fit: aov(formula = df$BMI ~ df$Income_Factor)
##
## $'df$Income_Factor'
##
                      {\tt diff}
                                  lwr
                                              upr
                                                      p adj
## Medium-Low -0.3649835 -0.9602329 0.2302659 0.3216591
               -1.6626948 -2.3901931 -0.9351966 0.0000003
## High-Low
## High-Medium -1.2977113 -1.9371019 -0.6583207 0.0000061
fit_all<-lm(Urine_Flow ~ BMI, data=df1)</pre>
fit_all
##
## Call:
## lm(formula = Urine_Flow ~ BMI, data = df1)
## Coefficients:
##
   (Intercept)
                         BMI
       1.44599
                    -0.01386
##
```

summary(fit_all)

```
##
## Call:
## lm(formula = Urine_Flow ~ BMI, data = df1)
##
## Residuals:
##
      Min
               1Q Median
                               ЗQ
                                      Max
## -1.1110 -0.5789 -0.3061 0.2275 11.4224
## Coefficients:
##
              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.445989 0.207435 6.971 1e-11 ***
              -0.013865 0.006902 -2.009 0.0451 *
## BMI
## ---
## Signif. codes: 0 '*** 0.001 '** 0.01 '* 0.05 '.' 0.1 ' 1
## Residual standard error: 1.097 on 498 degrees of freedom
## Multiple R-squared: 0.008038, Adjusted R-squared: 0.006046
## F-statistic: 4.035 on 1 and 498 DF, p-value: 0.0451
plot(c(min(df1$BMI,na.rm = TRUE),max(df1$BMI,na.rm = TRUE)),c(min(df1$Urine_Flow,na.rm = TRUE),max(df1$
    xlab="BMI [Kg/cm^2]",ylab="Urine Flow Rate [mL/sec]")
#ggplot(df1,aes(x=age,y=SBP,color=BMI_Factor)) +
#geom_smooth(method='lm', mapping =aes(x=age, y=SBP), se=FALSE)+
# geom jitter(width = .2)
points(df1$BMI,df1$Urine_Flow,pch=2,cex=0.6,col="red")
abline(fit_all,col="black")
```



```
fit_all<-lm(BMI ~ Income, data=df1)</pre>
fit_all
##
## Call:
## lm(formula = BMI ~ Income, data = df1)
## Coefficients:
## (Intercept)
                     Income
     3.138e+01
                 -3.796e-05
##
summary(fit_all)
##
## Call:
## lm(formula = BMI ~ Income, data = df1)
##
## Residuals:
##
       Min
                1Q Median
                                ЗQ
                                        Max
## -14.210 -4.939 -0.827
                             3.414 51.763
##
## Coefficients:
##
                 Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.138e+01 6.239e-01 50.301
               -3.796e-05 9.380e-06 -4.047
                                                 6e-05 ***
## Income
```

```
## ---
## Signif. codes: 0 '*** 0.001 '** 0.01 '* 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.009 on 498 degrees of freedom
## Multiple R-squared: 0.03185, Adjusted R-squared: 0.0299
## F-statistic: 16.38 on 1 and 498 DF, p-value: 6.004e-05

plot(c(min(df1$Income,na.rm = TRUE),max(df1$Income,na.rm = TRUE)),c(min(df1$BMI,na.rm = TRUE),max(df1$B\text{Xlab="Income} [$/yr]",ylab="Body Mass Index [Kg/cm^2]")
#ggplot(df1,aes(x=age,y=SBP,color=BMI_Factor)) +
#geom_smooth(method='lm',mapping =aes(x=age,y=SBP),se=FALSE)+
# geom_jitter(width = .2)
points(df1$Income,df1$BMI,pch=2,cex=0.6,col="red")
abline(fit_all,col="black")
```

