EPID674 Epidemiologic Data Analysis using R

Hypothesis Testing in R

Kelly Bakulski

Last compiled on November 03, 2020

## Chapter 4, Hypothesis Testing with R

# Set up options

# Install new packages

# Install packages. Do this only once.  
options(repos="https://cran.rstudio.com" )  
install.packages("epiDisplay")  
install.packages("Hmisc")  
install.packages("corrplot")  
  
# To avoid installing every time: change set up in curly brackets to eval=FALSE

# Load packages

# Specify file directories

# directory <- "/cloud/project/" #Class option when coding on RStudio Cloud  
  
# Three alternate options to customize and specify source data directory if programming on a personal computer (only need to do the one that works for you)  
# directory<-"M:/EPID674/Data/" # Alternate option 1: location of the parent dataset in afs space of school computer  
directory <- "/Volumes/GoogleDrive/My Drive/Teaching/EPID674/2020\_fall/EPID674\_Week4\_Class/" # Alternate option 2: location of data on my Macbook  
# directory<-"C:/Users/bakulski/Documents/EPID674/Data/" #Alternate option 3: location of data on a personal PC  
  
date <- format(Sys.Date(), "%Y%m%d")

# Load data, remake useful variables

# Check the file path  
paste0(directory, "nhanes3.rda")

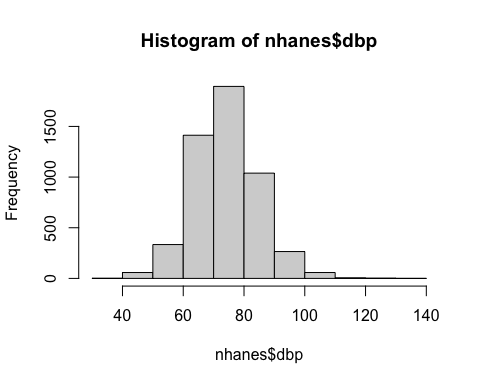
## [1] "/Volumes/GoogleDrive/My Drive/Teaching/EPID674/2020\_fall/EPID674\_Week4\_Class/nhanes3.rda"

# Load the saved R data  
load(paste0(directory, "nhanes3.rda"))  
### Or load RData if you saved the workspace image  
# load(".RData")  
  
  
# Remake a few variables from last class if they are no longer in your environment  
sex1 <- factor(nhanes$sex, levels = c(1, 2), labels = c("male", "female"))  
AGE5b <- cut(nhanes$age, quantile(nhanes$age, c(0, .2, .4, .6, .8, 1)), include.lowest = T) # quintiles  
AGE5c <- cut(nhanes$age, breaks = c(19, 40, 50, 60, 70, 90))  
age5c <- unclass(AGE5c)  
nhanes <- cbind(nhanes, sex1, AGE5b, AGE5c, age5c)

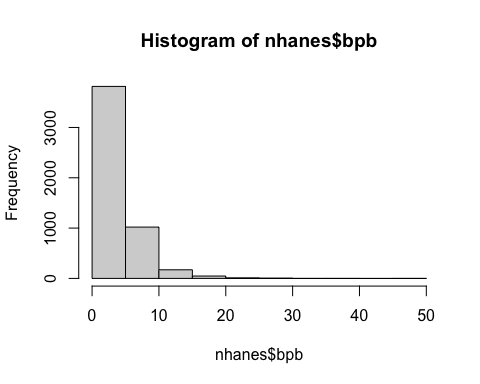
# Basic statistical tests in R

# Remember to check your distributions/assumptions

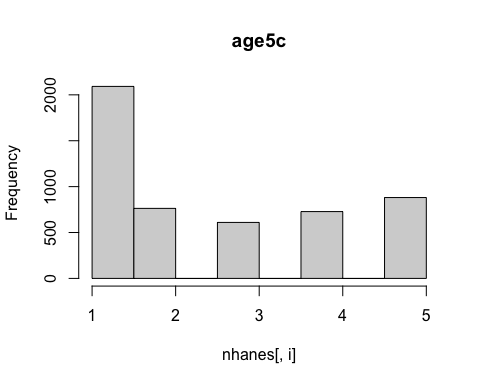
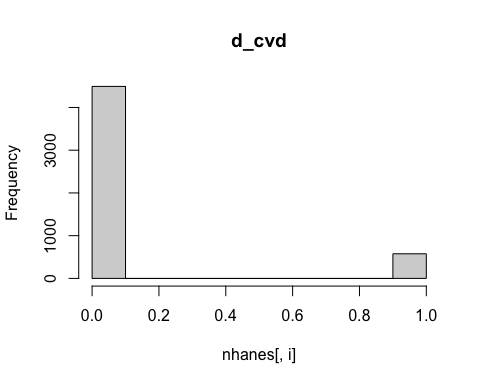
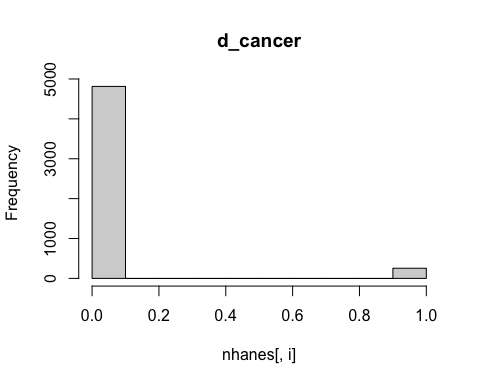
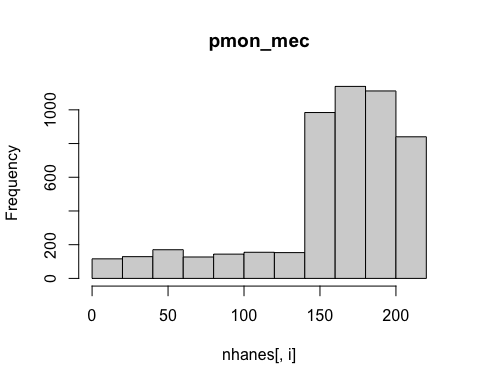
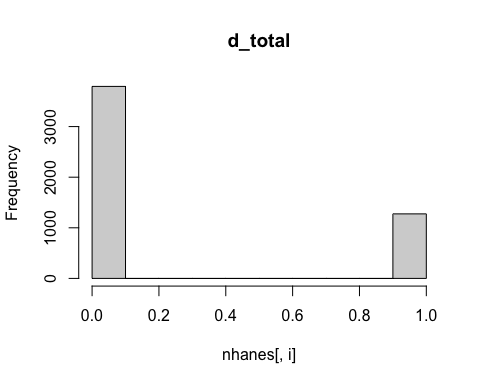
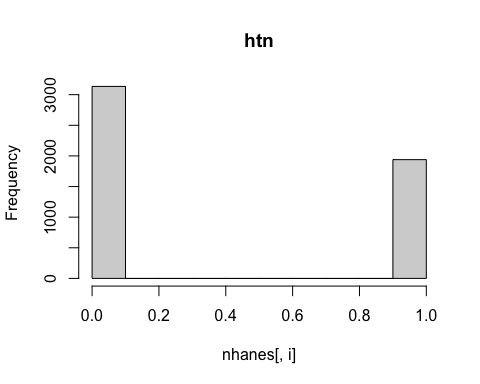
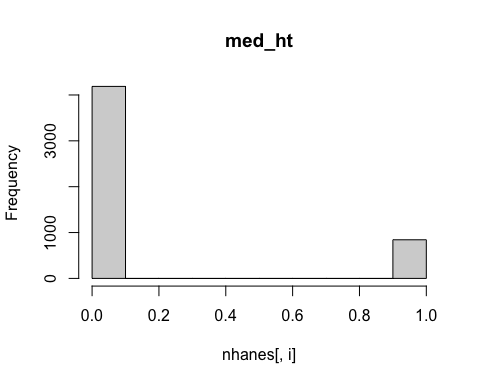
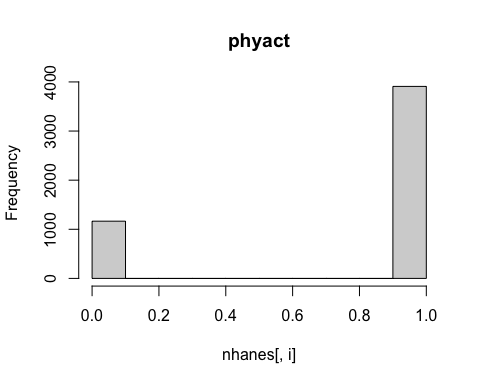
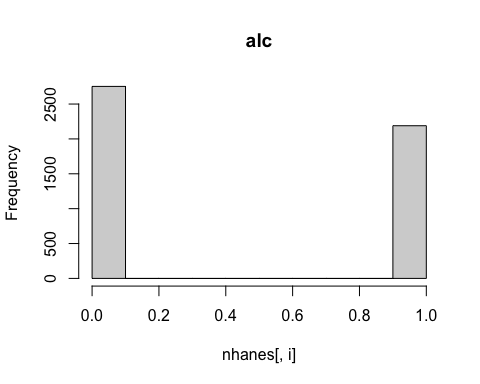
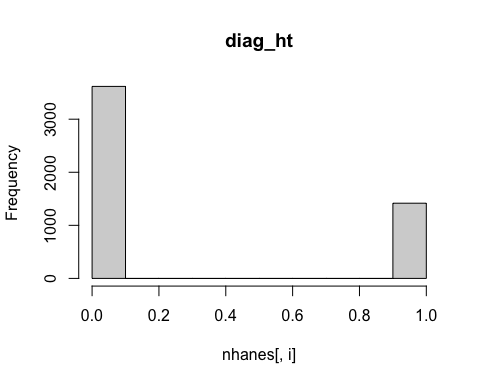
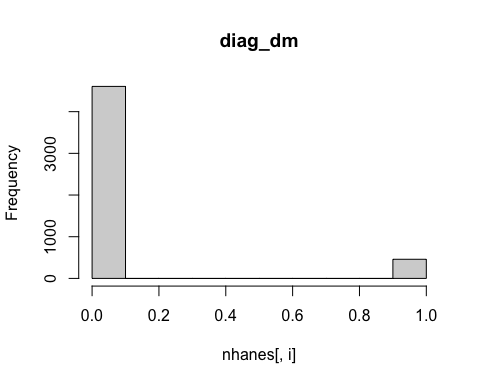
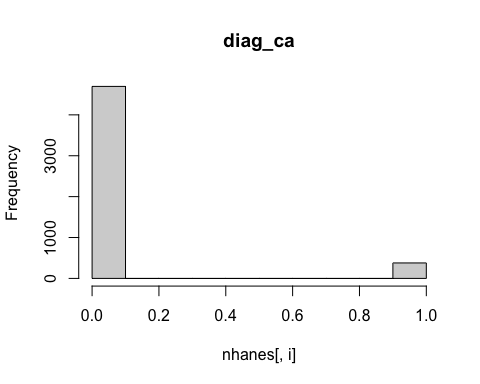
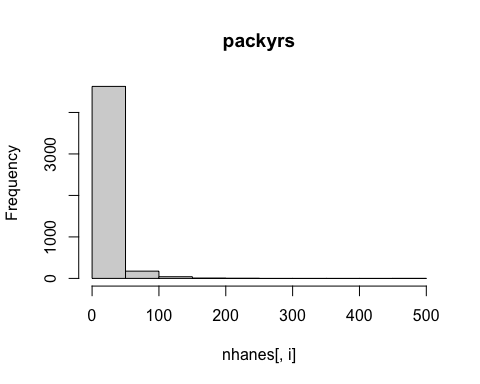
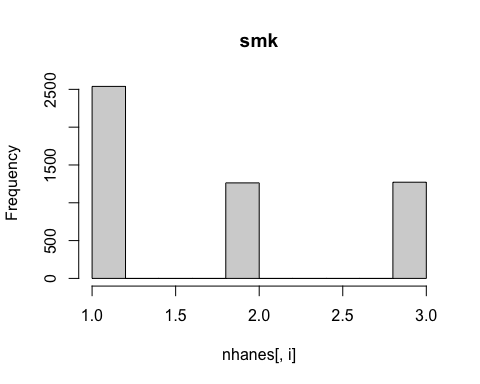
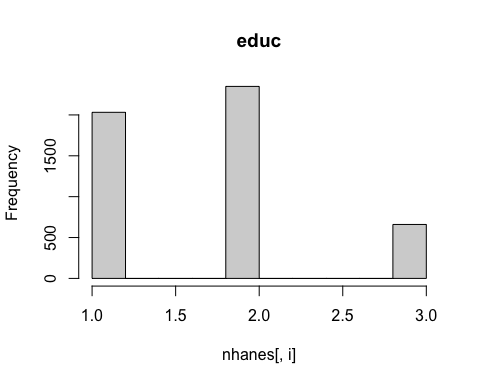
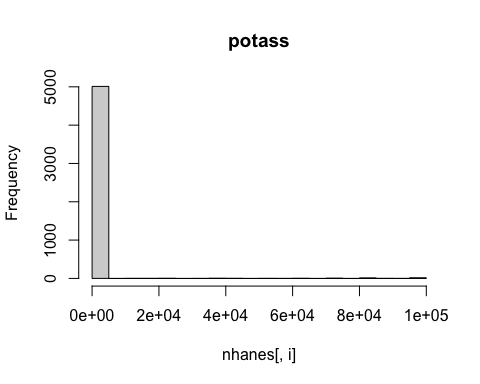
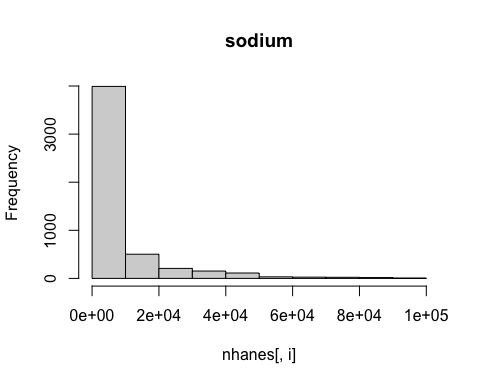
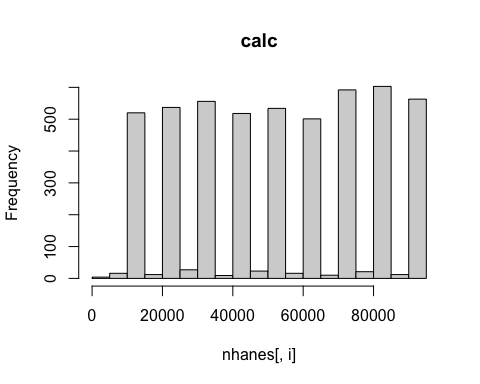
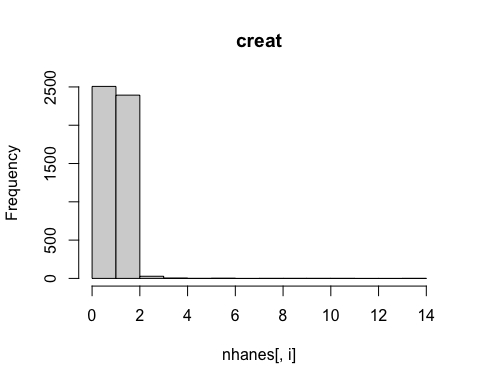
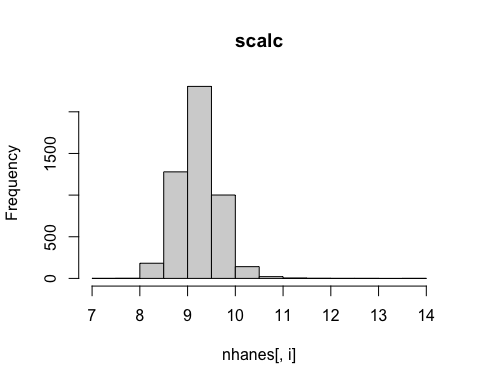
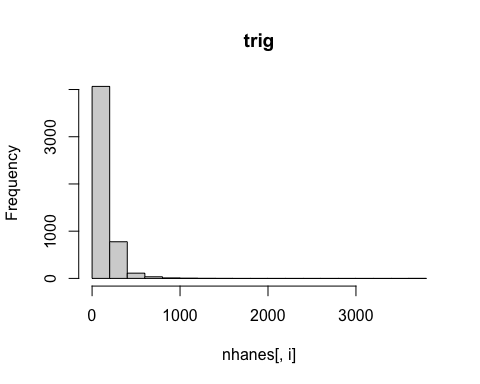
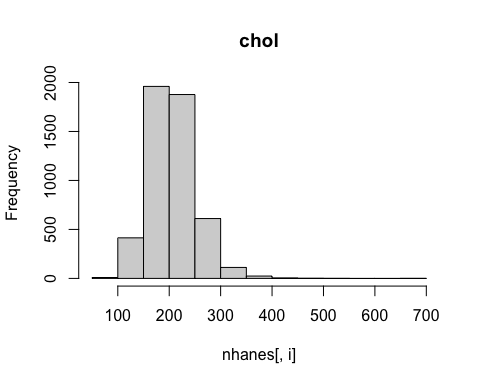
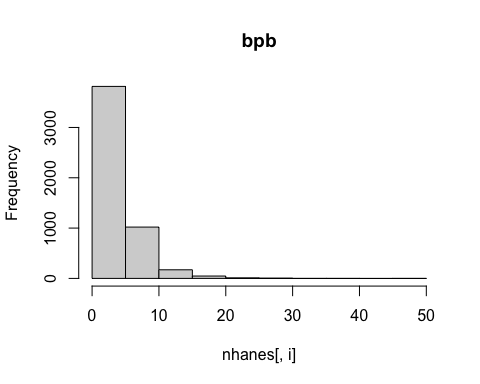
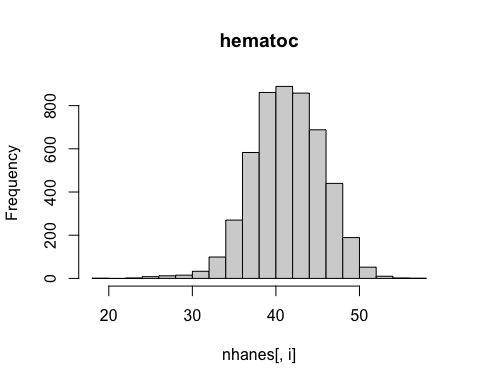
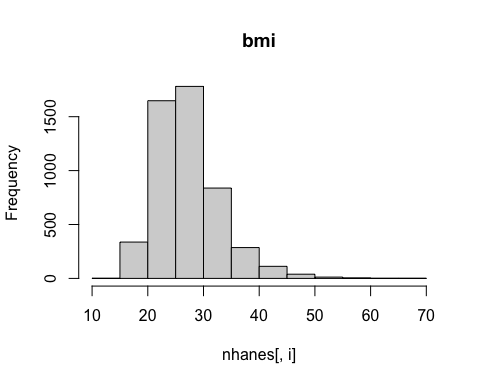
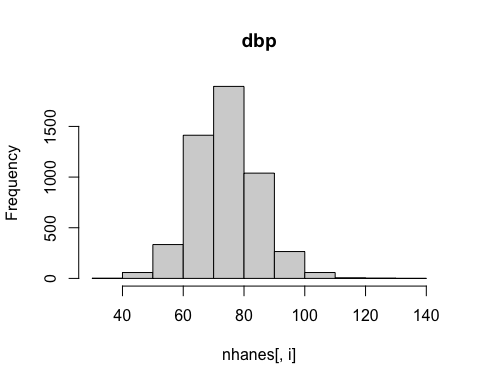
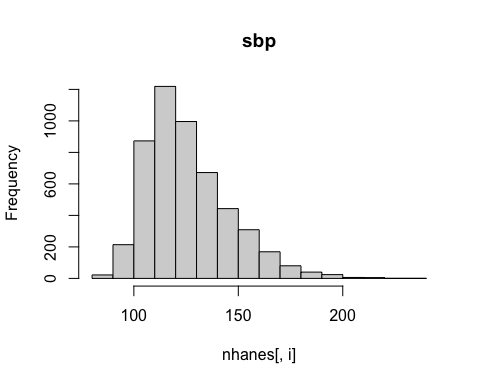
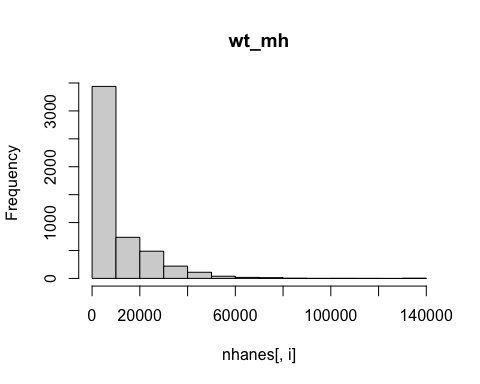
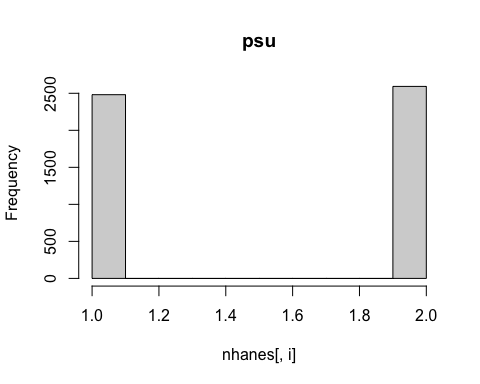
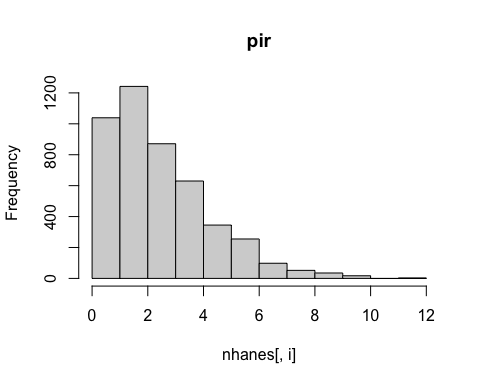
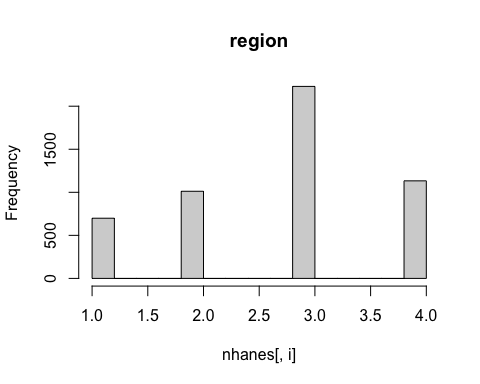
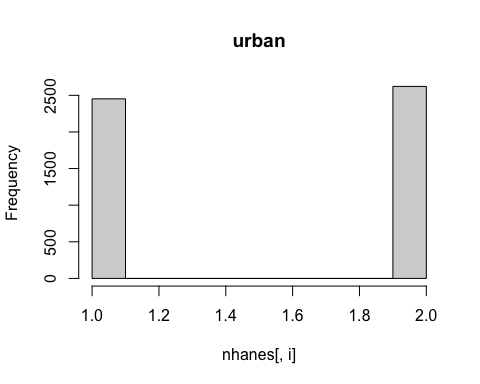
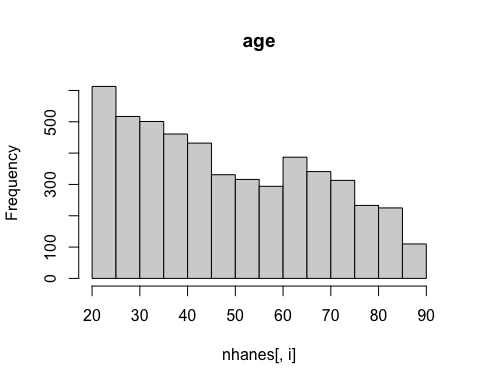
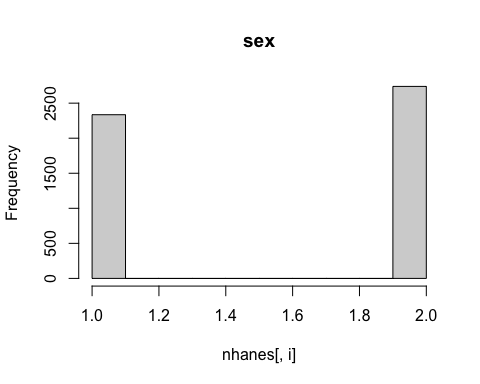
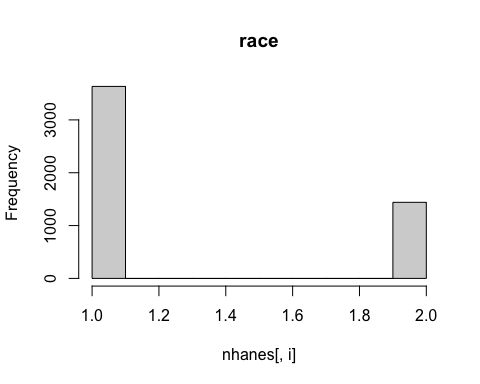
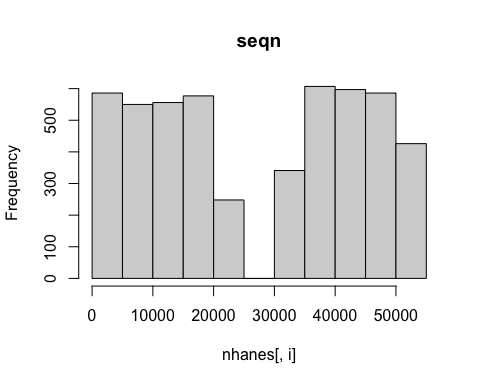
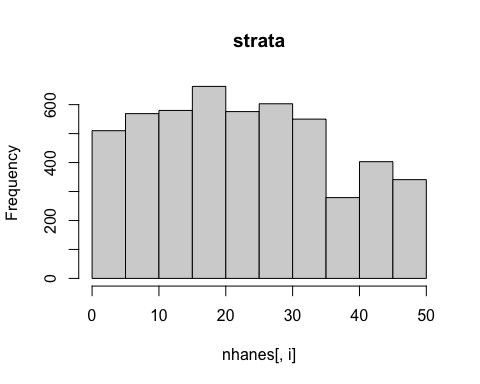
hist(nhanes$dbp)



hist(nhanes$bpb)



# for (i in 1:ncol(nhanes)) {  
# ifelse(is.numeric(nhanes[,i]), hist(nhanes[, i], main = colnames(nhanes)[i]), "non-numeric")  
# }  
#   
# for (i in 1:ncol(nhanes)) {  
# if(is.numeric(nhanes[,i])==T) {  
# hist(nhanes[, i], main = colnames(nhanes)[i])  
# } else { paste(colnames(nhanes[,i]),"is non-numeric")  
# }  
# }  
  
for (i in 1:ncol(nhanes)) {  
 if(is.numeric(nhanes[,i])==T) {  
 hist(nhanes[, i], main = colnames(nhanes)[i])  
 }  
}



# Correlation tests

cor.test(nhanes$age, nhanes$bmi)

##   
## Pearson's product-moment correlation  
##   
## data: nhanes$age and nhanes$bmi  
## t = 2.7253, df = 5062, p-value = 0.006446  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.01074507 0.06575154  
## sample estimates:  
## cor   
## 0.0382773

cor.test(nhanes$age, nhanes$bpb, method = "spearman")

## Warning in cor.test.default(nhanes$age, nhanes$bpb, method = "spearman"): Cannot  
## compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: nhanes$age and nhanes$bpb  
## S = 1.5977e+10, p-value < 2.2e-16  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.2661545

# Correlation matrices

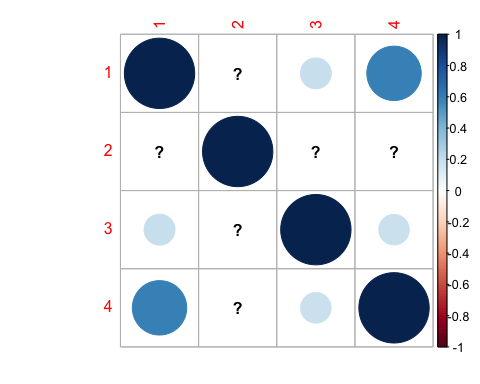
nhanes.d <- cbind(nhanes$age, nhanes$bmi, nhanes$bpb, nhanes$sbp)  
rcorr(nhanes.d)

## [,1] [,2] [,3] [,4]  
## [1,] 1.00 0.04 0.19 0.60  
## [2,] 0.04 1.00 -0.05 0.17  
## [3,] 0.19 -0.05 1.00 0.19  
## [4,] 0.60 0.17 0.19 1.00  
##   
## n  
## [,1] [,2] [,3] [,4]  
## [1,] 5074 5064 5074 5074  
## [2,] 5064 5064 5064 5064  
## [3,] 5074 5064 5074 5074  
## [4,] 5074 5064 5074 5074  
##   
## P  
## [,1] [,2] [,3] [,4]   
## [1,] 0.0064 0.0000 0.0000  
## [2,] 0.0064 0.0001 0.0000  
## [3,] 0.0000 0.0001 0.0000  
## [4,] 0.0000 0.0000 0.0000

rcorr(nhanes.d, type = "spearman")

## [,1] [,2] [,3] [,4]  
## [1,] 1.00 0.11 0.27 0.61  
## [2,] 0.11 1.00 -0.01 0.24  
## [3,] 0.27 -0.01 1.00 0.27  
## [4,] 0.61 0.24 0.27 1.00  
##   
## n  
## [,1] [,2] [,3] [,4]  
## [1,] 5074 5064 5074 5074  
## [2,] 5064 5064 5064 5064  
## [3,] 5074 5064 5074 5074  
## [4,] 5074 5064 5074 5074  
##   
## P  
## [,1] [,2] [,3] [,4]   
## [1,] 0.0000 0.0000 0.0000  
## [2,] 0.0000 0.5264 0.0000  
## [3,] 0.0000 0.5264 0.0000  
## [4,] 0.0000 0.0000 0.0000

corrplot(cor(nhanes.d))



nhanes.e <- na.omit(nhanes[, c("age", "bmi", "bpb", "sbp")])  
cor(nhanes.e)

## age bmi bpb sbp  
## age 1.0000000 0.03827730 0.19238154 0.5982064  
## bmi 0.0382773 1.00000000 -0.05346208 0.1693454  
## bpb 0.1923815 -0.05346208 1.00000000 0.1867691  
## sbp 0.5982064 0.16934537 0.18676909 1.0000000

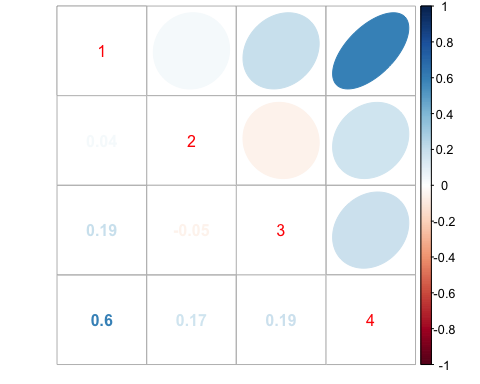
nhanes.f <- cbind(nhanes.e$age, nhanes.e$bmi, nhanes.e$bpb, nhanes.e$sbp)  
rcorr(nhanes.f)

## [,1] [,2] [,3] [,4]  
## [1,] 1.00 0.04 0.19 0.60  
## [2,] 0.04 1.00 -0.05 0.17  
## [3,] 0.19 -0.05 1.00 0.19  
## [4,] 0.60 0.17 0.19 1.00  
##   
## n= 5064   
##   
##   
## P  
## [,1] [,2] [,3] [,4]   
## [1,] 0.0064 0.0000 0.0000  
## [2,] 0.0064 0.0001 0.0000  
## [3,] 0.0000 0.0001 0.0000  
## [4,] 0.0000 0.0000 0.0000

rcorr(nhanes.f, type = "spearman")

## [,1] [,2] [,3] [,4]  
## [1,] 1.00 0.11 0.27 0.61  
## [2,] 0.11 1.00 -0.01 0.24  
## [3,] 0.27 -0.01 1.00 0.27  
## [4,] 0.61 0.24 0.27 1.00  
##   
## n= 5064   
##   
##   
## P  
## [,1] [,2] [,3] [,4]   
## [1,] 0.0000 0.0000 0.0000  
## [2,] 0.0000 0.5264 0.0000  
## [3,] 0.0000 0.5264 0.0000  
## [4,] 0.0000 0.0000 0.0000

corrplot.mixed(cor(nhanes.f), upper = "ellipse", lower = "number")



# T-test and Wilcoxon test

t.test(nhanes$dbp ~ nhanes$sex1)

##   
## Welch Two Sample t-test  
##   
## data: nhanes$dbp by nhanes$sex1  
## t = 15.932, df = 4909.1, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 4.066115 5.207214  
## sample estimates:  
## mean in group male mean in group female   
## 77.02184 72.38518

wilcox.test(nhanes$bpb ~ nhanes$sex1)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: nhanes$bpb by nhanes$sex1  
## W = 4595584, p-value < 2.2e-16  
## alternative hypothesis: true location shift is not equal to 0

t.test(log(nhanes$bpb) ~ nhanes$sex1, var.equal = T)

##   
## Two Sample t-test  
##   
## data: log(nhanes$bpb) by nhanes$sex1  
## t = 29.104, df = 5072, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.5259266 0.6018957  
## sample estimates:  
## mean in group male mean in group female   
## 1.4132205 0.8493094

# ANOVA

tapply(log(nhanes$bpb), nhanes$AGE5b, mean)

## [20,30] (30,40] (40,54] (54,69] (69,90]   
## 0.8485350 0.9249257 1.1524859 1.3121878 1.3380641

aov(log(nhanes$bpb) ~ nhanes$AGE5b)

## Call:  
## aov(formula = log(nhanes$bpb) ~ nhanes$AGE5b)  
##   
## Terms:  
## nhanes$AGE5b Residuals  
## Sum of Squares 203.5226 2597.3352  
## Deg. of Freedom 4 5069  
##   
## Residual standard error: 0.7158184  
## Estimated effects may be unbalanced

bpb.aov <- aov(log(nhanes$bpb) ~ nhanes$AGE5b)  
anova(bpb.aov)

## Analysis of Variance Table  
##   
## Response: log(nhanes$bpb)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## nhanes$AGE5b 4 203.52 50.881 99.299 < 2.2e-16 \*\*\*  
## Residuals 5069 2597.34 0.512   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(lm(log(nhanes$bpb) ~ nhanes$AGE5b))

## Analysis of Variance Table  
##   
## Response: log(nhanes$bpb)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## nhanes$AGE5b 4 203.52 50.881 99.299 < 2.2e-16 \*\*\*  
## Residuals 5069 2597.34 0.512   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

pairwise.t.test(log(nhanes$bpb), nhanes$AGE5b)

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: log(nhanes$bpb) and nhanes$AGE5b   
##   
## [20,30] (30,40] (40,54] (54,69]  
## (30,40] 0.03 - - -   
## (40,54] < 2e-16 1.1e-11 - -   
## (54,69] < 2e-16 < 2e-16 1.5e-06 -   
## (69,90] < 2e-16 < 2e-16 4.4e-08 0.42   
##   
## P value adjustment method: holm

pairwise.t.test(log(nhanes$bpb), nhanes$AGE5b, p.adj = "bonf")

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: log(nhanes$bpb) and nhanes$AGE5b   
##   
## [20,30] (30,40] (40,54] (54,69]  
## (30,40] 0.15 - - -   
## (40,54] < 2e-16 2.1e-11 - -   
## (54,69] < 2e-16 < 2e-16 5.1e-06 -   
## (69,90] < 2e-16 < 2e-16 1.1e-07 1.00   
##   
## P value adjustment method: bonferroni

# Chi-square and Fisher’s tests

chisq.test(nhanes$sex, nhanes$smk)

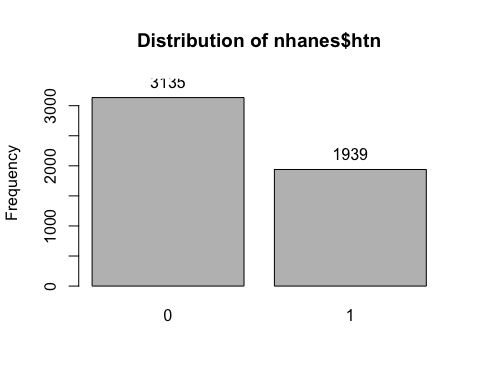
##   
## Pearson's Chi-squared test  
##   
## data: nhanes$sex and nhanes$smk  
## X-squared = 351.08, df = 2, p-value < 2.2e-16

fisher.test(nhanes$race, nhanes$htn)

##   
## Fisher's Exact Test for Count Data  
##   
## data: nhanes$race and nhanes$htn  
## p-value = 0.04354  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 1.002389 1.291976  
## sample estimates:  
## odds ratio   
## 1.138167

# Odds Ratios

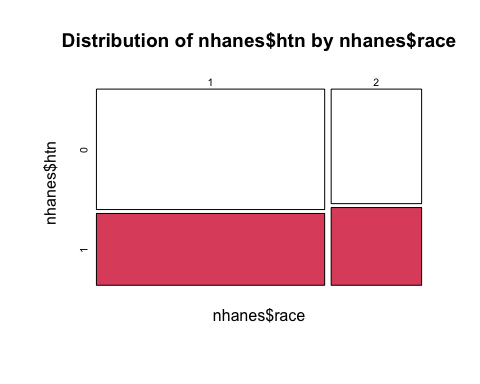
tab1(nhanes$htn)



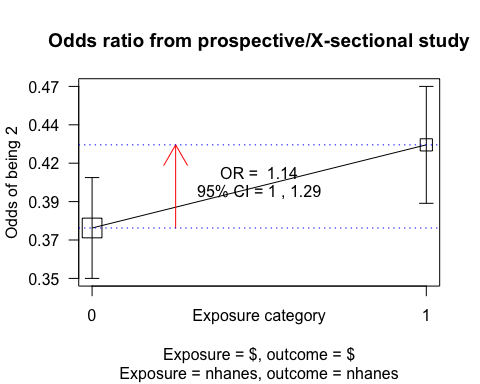
## nhanes$htn :   
## Frequency Percent Cum. percent  
## 0 3135 61.8 61.8  
## 1 1939 38.2 100.0  
## Total 5074 100.0 100.0

tabpct(nhanes$race, nhanes$htn)

##   
## Original table   
## nhanes$htn  
## nhanes$race 0 1 Total  
## 1 2277 1357 3634  
## 2 858 582 1440  
## Total 3135 1939 5074  
##   
## Row percent   
## nhanes$htn  
## nhanes$race 0 1 Total  
## 1 2277 1357 3634  
## (62.7) (37.3) (100)  
## 2 858 582 1440  
## (59.6) (40.4) (100)  
##   
## Column percent   
## nhanes$htn  
## nhanes$race 0 % 1 %  
## 1 2277 (72.6) 1357 (70)  
## 2 858 (27.4) 582 (30)  
## Total 3135 (100) 1939 (100)



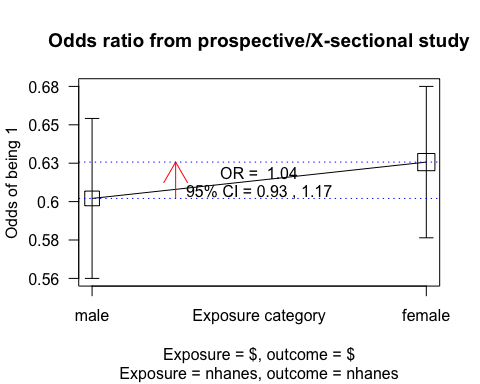
cc(nhanes$race, nhanes$htn)



##   
## nhanes$htn  
## nhanes$race 0 1 Total  
## 1 2277 1357 3634  
## 2 858 582 1440  
## Total 3135 1939 5074  
##   
## OR = 1.14   
## 95% CI = 1, 1.29   
## Chi-squared = 4.13, 1 d.f., P value = 0.042  
## Fisher's exact test (2-sided) P value = 0.044

# Mantel-Haenszel Odds Ratio

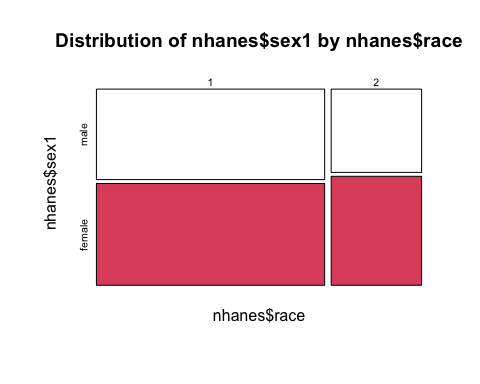
cc(nhanes$htn, nhanes$sex1)



##   
## nhanes$sex1  
## nhanes$htn male female Total  
## 0 1454 1681 3135  
## 1 881 1058 1939  
## Total 2335 2739 5074  
##   
## OR = 1.04   
## 95% CI = 0.93, 1.16   
## Chi-squared = 0.43, 1 d.f., P value = 0.512  
## Fisher's exact test (2-sided) P value = 0.524

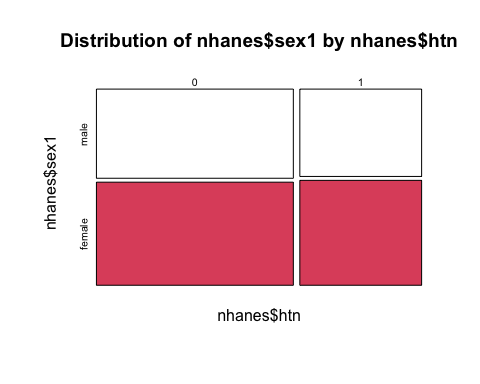
tabpct(nhanes$race, nhanes$sex1)

##   
## Original table   
## nhanes$sex1  
## nhanes$race male female Total  
## 1 1711 1923 3634  
## 2 624 816 1440  
## Total 2335 2739 5074  
##   
## Row percent   
## nhanes$sex1  
## nhanes$race male female Total  
## 1 1711 1923 3634  
## (47.1) (52.9) (100)  
## 2 624 816 1440  
## (43.3) (56.7) (100)  
##   
## Column percent   
## nhanes$sex1  
## nhanes$race male % female %  
## 1 1711 (73.3) 1923 (70.2)  
## 2 624 (26.7) 816 (29.8)  
## Total 2335 (100) 2739 (100)



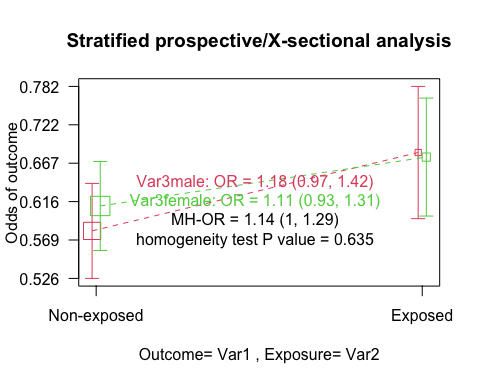
tabpct(nhanes$htn, nhanes$sex1)

##   
## Original table   
## nhanes$sex1  
## nhanes$htn male female Total  
## 0 1454 1681 3135  
## 1 881 1058 1939  
## Total 2335 2739 5074  
##   
## Row percent   
## nhanes$sex1  
## nhanes$htn male female Total  
## 0 1454 1681 3135  
## (46.4) (53.6) (100)  
## 1 881 1058 1939  
## (45.4) (54.6) (100)  
##   
## Column percent   
## nhanes$sex1  
## nhanes$htn male % female %  
## 0 1454 (62.3) 1681 (61.4)  
## 1 881 (37.7) 1058 (38.6)  
## Total 2335 (100) 2739 (100)



mhor(nhanes$htn, nhanes$race, nhanes$sex1)

##   
## Stratified analysis by Var3   
## OR lower lim. upper lim. P value  
## Var3 male 1.18 0.970 1.42 0.0916  
## Var3 female 1.11 0.932 1.31 0.2467  
## M-H combined 1.14 1.003 1.29 0.0444  
##   
## M-H Chi2(1) = 4.04 , P value = 0.044   
## Homogeneity test, chi-squared 1 d.f. = 0.23 , P value = 0.635



# Do Exercise 4