EPID674 Epidemiologic Data Analysis using R

Homework 4 Answer Key

Kelly Bakulski

Last compiled on November 18, 2020

# Setup libraries and dataset

knitr::opts\_chunk$set(echo = TRUE)  
  
library(foreign)  
library(epiDisplay)

## Loading required package: survival

## Loading required package: MASS

## Loading required package: nnet

library(stats)  
library(Hmisc)

## Loading required package: lattice

##   
## Attaching package: 'lattice'

## The following object is masked from 'package:epiDisplay':  
##   
## dotplot

## Loading required package: Formula

## Loading required package: ggplot2

##   
## Attaching package: 'ggplot2'

## The following object is masked from 'package:epiDisplay':  
##   
## alpha

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

library(gmodels)

##   
## Attaching package: 'gmodels'

## The following object is masked from 'package:epiDisplay':  
##   
## ci

#directory <- "/cloud/project/"  
directory <- "~/Google Drive File Stream/My Drive/Teaching/EPID674/2020\_fall/EPID674\_Week4\_Class/"  
load(paste0(directory, "bpa.rda"))

## Using the dataset “bpa.rda”, answer the following:

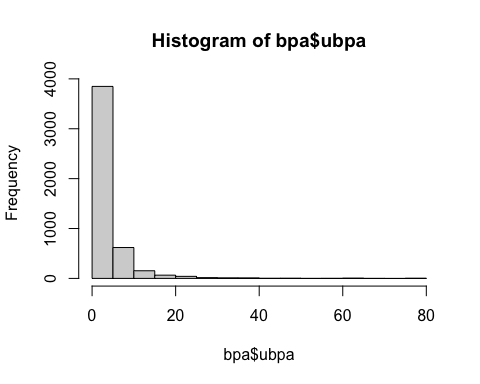
# Problem 1.

## Does urinary BPA differ by age group (20-39, 40-59, 60+)?

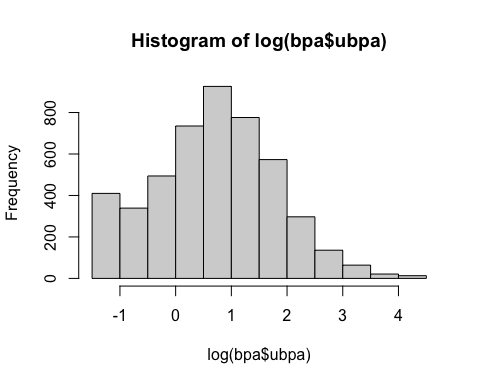
## Age group  
age3<-cut(bpa$age, breaks=c(20,39,59,90), include.lowest=T)  
summary(age3)

## [20,39] (39,59] (59,90]   
## 1693 1474 1617

#check normality of ubpa  
hist(bpa$ubpa) #UPBA variable is highly right skewed. Need to log transform UBPA or perform non-parametric test.



hist(log(bpa$ubpa))



tapply(log(bpa$ubpa), age3, mean) #set expectation

## [20,39] (39,59] (59,90]   
## 0.9187357 0.6735022 0.5170613

anova(lm(log(bpa$ubpa)~age3))

## Analysis of Variance Table  
##   
## Response: log(bpa$ubpa)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## age3 2 135.9 67.945 58.984 < 2.2e-16 \*\*\*  
## Residuals 4781 5507.3 1.152   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

kruskal.test(bpa$ubpa~age3)

##   
## Kruskal-Wallis rank sum test  
##   
## data: bpa$ubpa by age3  
## Kruskal-Wallis chi-squared = 122.02, df = 2, p-value < 2.2e-16

summary(lm(log(bpa$ubpa)~age3))

##   
## Call:  
## lm(formula = log(bpa$ubpa) ~ age3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.1917 -0.7364 0.0196 0.7067 3.8625   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.91874 0.02608 35.221 < 2e-16 \*\*\*  
## age3(39,59] -0.24523 0.03823 -6.414 1.56e-10 \*\*\*  
## age3(59,90] -0.40167 0.03732 -10.763 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.073 on 4781 degrees of freedom  
## Multiple R-squared: 0.02408, Adjusted R-squared: 0.02367   
## F-statistic: 58.98 on 2 and 4781 DF, p-value: < 2.2e-16

## To check linear trend, the factor variable age3 should be converted to a numeric variable  
age3n<-unclass(age3)  
tapply(log(bpa$ubpa), age3n, mean) #set expectation

## 1 2 3   
## 0.9187357 0.6735022 0.5170613

summary(lm(log(bpa$ubpa)~age3n))

##   
## Call:  
## lm(formula = log(bpa$ubpa) ~ age3n)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.1783 -0.7230 0.0276 0.6909 3.8765   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.10652 0.04014 27.56 <2e-16 \*\*\*  
## age3n -0.20115 0.01866 -10.78 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.073 on 4782 degrees of freedom  
## Multiple R-squared: 0.02372, Adjusted R-squared: 0.02352   
## F-statistic: 116.2 on 1 and 4782 DF, p-value: < 2.2e-16

# Yes, urinary bpa levels differ by age group. Levels are higher in lower ages.

# Problem 2

## Does urinary BPA differ by geneder?

tapply(log(bpa$ubpa), bpa$gender, mean) #set expectation

## 1 2   
## 0.7780433 0.6405700

t.test(log(bpa$ubpa)~bpa$gender) #perform parametric test

##   
## Welch Two Sample t-test  
##   
## data: log(bpa$ubpa) by bpa$gender  
## t = 4.3921, df = 4780.9, p-value = 1.147e-05  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.07611069 0.19883591  
## sample estimates:  
## mean in group 1 mean in group 2   
## 0.7780433 0.6405700

t.test(log(bpa$ubpa)~bpa$gender, var.equal=T) #another option

##   
## Two Sample t-test  
##   
## data: log(bpa$ubpa) by bpa$gender  
## t = 4.3836, df = 4782, p-value = 1.192e-05  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.07599152 0.19895507  
## sample estimates:  
## mean in group 1 mean in group 2   
## 0.7780433 0.6405700

wilcox.test(bpa$ubpa~bpa$gender) #non-parametric option

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: bpa$ubpa by bpa$gender  
## W = 3062316, p-value = 1.977e-05  
## alternative hypothesis: true location shift is not equal to 0

# Yes, urinary bpa levels differ by gender. They are higher in gender 1 and lower in gender 2.

# Problem 3

## Does urinary BPA differ by smoking status?

#watch out for NaN. Need to recode as missing  
table(bpa$SMK, useNA = "always")

##   
## 0 1 2 NaN <NA>   
## 2490 1222 1067 5 0

bpa$SMK[is.nan(bpa$SMK)]<-NA #Assign value of NA to any entries that are non at number "NaN"  
table(bpa$SMK, useNA = "always")

##   
## 0 1 2 <NA>   
## 2490 1222 1067 5

tapply(log(bpa$ubpa), bpa$SMK, mean) #Set expectation

## 0 1 2   
## 0.7033241 0.6468335 0.7848018

kruskal.test(bpa$ubpa~bpa$SMK)

##   
## Kruskal-Wallis rank sum test  
##   
## data: bpa$ubpa by bpa$SMK  
## Kruskal-Wallis chi-squared = 10.599, df = 2, p-value = 0.004993

pairwise.t.test(log(bpa$ubpa),bpa$SMK)

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: log(bpa$ubpa) and bpa$SMK   
##   
## 0 1   
## 1 0.1364 -   
## 2 0.0807 0.0073  
##   
## P value adjustment method: holm

anova(lm(log(bpa$ubpa)~factor(bpa$SMK)))

## Analysis of Variance Table  
##   
## Response: log(bpa$ubpa)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(bpa$SMK) 2 10.9 5.458 4.6293 0.009806 \*\*  
## Residuals 4776 5631.0 1.179   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(lm(log(bpa$ubpa)~factor(bpa$SMK)))

##   
## Call:  
## lm(formula = log(bpa$ubpa) ~ factor(bpa$SMK))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.0578 -0.7033 0.0386 0.7141 3.6762   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.70332 0.02176 32.322 <2e-16 \*\*\*  
## factor(bpa$SMK)1 -0.05649 0.03793 -1.490 0.1364   
## factor(bpa$SMK)2 0.08148 0.03973 2.051 0.0403 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.086 on 4776 degrees of freedom  
## (5 observations deleted due to missingness)  
## Multiple R-squared: 0.001935, Adjusted R-squared: 0.001517   
## F-statistic: 4.629 on 2 and 4776 DF, p-value: 0.009806

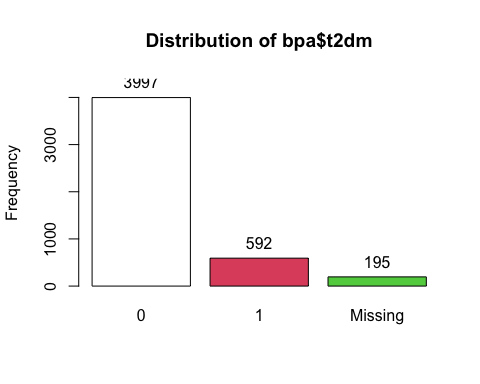
# Yes, urinary BPA differs by smoking status. The biggest difference is between smoking group 1 and smoking group 2.

# Problem 4

## T2DM can be defined as hemoglobin A1c ≥ 6.5% OR use of diabetes medication. Note: participants with high A1c and taking medication should also be considered T2DM cases.

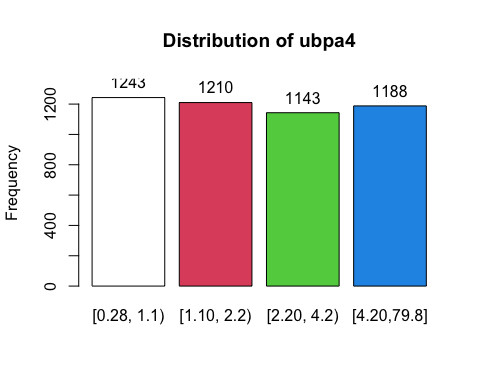
## First, create a binary variable for T2DM using a1c and dmmed. How many T2DM cases are in each quartile of urinary BPA?

## we already created this in the previous exercise  
bpa$t2dm<-ifelse(bpa$a1c>=6.5|bpa$dmmed==1, 1, 0)  
tab1(bpa$t2dm)



## bpa$t2dm :   
## Frequency %(NA+) %(NA-)  
## 0 3997 83.5 87.1  
## 1 592 12.4 12.9  
## <NA> 195 4.1 0.0  
## Total 4784 100.0 100.0

ubpa4<-cut2(bpa$ubpa, g=4)  
tab1(ubpa4)



## ubpa4 :   
## Frequency Percent Cum. percent  
## [0.28, 1.1) 1243 26.0 26.0  
## [1.10, 2.2) 1210 25.3 51.3  
## [2.20, 4.2) 1143 23.9 75.2  
## [4.20,79.8] 1188 24.8 100.0  
## Total 4784 100.0 100.0

table(ubpa4, bpa$t2dm)

##   
## ubpa4 0 1  
## [0.28, 1.1) 1044 154  
## [1.10, 2.2) 1001 152  
## [2.20, 4.2) 950 142  
## [4.20,79.8] 1002 144

# or  
ubpa4<-cut(bpa$ubpa, quantile(bpa$ubpa, c(0,0.25,0.5,0.75,1)), include.lowest=T)  
summary(ubpa4)

## [0.28,1] (1,2.1] (2.1,4.1] (4.1,79.8]   
## 1243 1210 1143 1188

tapply(ubpa4, bpa$t2dm, summary)

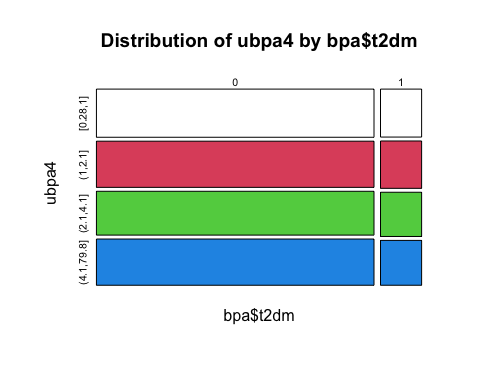
## $`0`  
## [0.28,1] (1,2.1] (2.1,4.1] (4.1,79.8]   
## 1044 1001 950 1002   
##   
## $`1`  
## [0.28,1] (1,2.1] (2.1,4.1] (4.1,79.8]   
## 154 152 142 144

# There are 154 T2DM cases in the lowest quartile of urinary bpa, 152 cases in the second quartile, 142 cases in the thir quartile, and 144 cases in the highest quartile.

##Compute the odds ratios of T2DM for each of the upper three quartiles of urinary BPA, as compared with the lowest quartile of urinary BPA.

tabpct(bpa$t2dm,ubpa4)

##   
## Original table   
## ubpa4  
## bpa$t2dm [0.28,1] (1,2.1] (2.1,4.1] (4.1,79.8] Total  
## 0 1044 1001 950 1002 3997  
## 1 154 152 142 144 592  
## Total 1198 1153 1092 1146 4589  
##   
## Row percent   
## ubpa4  
## bpa$t2dm [0.28,1] (1,2.1] (2.1,4.1] (4.1,79.8] Total  
## 0 1044 1001 950 1002 3997  
## (26.1) (25) (23.8) (25.1) (100)  
## 1 154 152 142 144 592  
## (26) (25.7) (24) (24.3) (100)  
##   
## Column percent   
## ubpa4  
## bpa$t2dm [0.28,1] % (1,2.1] % (2.1,4.1] % (4.1,79.8]  
## 0 1044 (87.1) 1001 (86.8) 950 (87) 1002  
## 1 154 (12.9) 152 (13.2) 142 (13) 144  
## Total 1198 (100) 1153 (100) 1092 (100) 1146  
## ubpa4  
## bpa$t2dm %  
## 0 (87.4)  
## 1 (12.6)  
## Total (100)

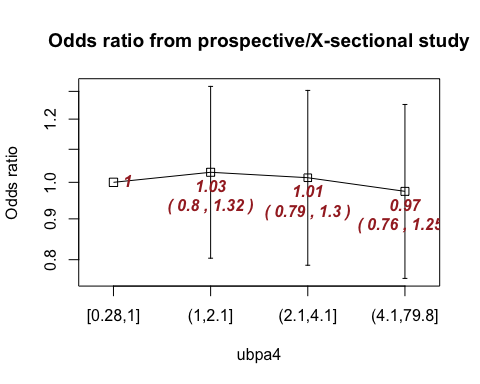


chisq.test(bpa$t2dm,ubpa4)

##   
## Pearson's Chi-squared test  
##   
## data: bpa$t2dm and ubpa4  
## X-squared = 0.20897, df = 3, p-value = 0.9761

cc(bpa$t2dm,ubpa4)

##   
## ubpa4  
## bpa$t2dm [0.28,1] (1,2.1] (2.1,4.1] (4.1,79.8]  
## 0 1044 1001 950 1002   
## 1 154 152 142 144   
##   
## Odds ratio 1 1.03 1.01 0.97   
## lower 95% CI 0.8 0.79 0.76   
## upper 95% CI 1.32 1.3 1.25   
##   
## Chi-squared = 0.209 , 3 d.f., P value = 0.976   
## Fisher's exact test (2-sided) P value = 0.976

 # There is no linear trend across the urinary BPA quartiles as shown in the plot above. The prevalence of T2DM is NOT associated with urinary BPA.