Assignment 4 solutions

BIOSTAT 213: Introduction to Computing in the R Software Environment

Due: Wednesday, November 27, 2019

Collaboration with other students in this course is permitted on this assignment. However, you must write all work—including R code—in your own voice, as much as possible, and acknowledge your collaborators as thoroughly as possible.

The assignment is due online, in the Collaborative Learning Environment, by 11:59 pm, Wednesday, November 27, 2019. You must submit a Word-compatible document or a PDF file.

For each of the below, share your code as well as the output (if there is any output). If your answer to a question is unclear from your R output, or difficult to identify, include an answer in text form (in addition to the code and output).

1. As you may have noticed, the `round()` function eliminates any trailing 0s. For example:

```
> round(9.403,2)
[1] 9.4
```

   One way around this is to use the `format()` function:

```
> format(9.403,nsmall=2,digits=2)
[1] "9.40"
> round(9.403,2)
[1] 9.4
```

   Write a rounding function that accepts 2 arguments: the number that should be rounded and the number of digits. Here's an example function:

```
> Round(3.403,2)
[1] "3.40"
> Round(pi,4)
[1] "3.1416"
> Round(-90.0004,3)
[1] "-90.000"
```

   Include some tests of the function in your response. You may use the inputs above if you wish.

   (Reminder: as a general rule in R and other statistical software, rounding functions should only be used when presenting final numbers. In other words, do not round in the "middle" of your work. It will lead to imprecision.)

```
> Round<-function(x,digits){format(x,nsmall=digits,digits=digits)}
> Round(3.403,2)
[1] "3.40"
> Round(pi,4)
[1] "3.1416"
> Round(-90.0004,3)
[1] "-90.000"
```

2. Write a function that will implement a two-sample t-test, using `t.test()`. The function should output a formatted version of the results, showing just the difference in means and the associated 95% confidence interval. Note that you should take the difference by subtracting the mean in the second group from the mean in the first group. Optionally, you may consider including the custom rounding function from above. Here's an example function:

```
> difference(y=ii$bmxbmi,x=ii$gender)
[1] "1.23 (95% CI:0.79, 1.67)"
> difference(y=ii$bpxsy1,x=ii$hs)
[1] "4.30 (95% CI:3.22, 5.37)"
```

Include some tests of the function in your response. You may use the inputs above if you wish.

To be clear: you can access these numbers from `t.test()`. The group means are accessible via `$estimate` while the confidence interval is accessible via `$conf.int`.

```
> difference<-function(y,x){
+   tt<-t.test(y~x)
+   pp<-Round(as.numeric(tt$estimate[1]-tt$estimate[2]),2)
+   ii<-paste(Round(tt$conf.int,2),collapse=', ')
+   paste(pp,' (95% CI:',ii,')',sep='')
+ }
> difference(y=ii$bmxbmi,x=ii$gender)
[1] "1.23 (95% CI:0.79, 1.67)"
> difference(y=ii$bpxsy1,x=ii$hs)
[1] "4.30 (95% CI:3.22, 5.37)"
```

3. Write a function that will perform a bivariate test on categorical data. The function should accept 3 arguments: 2 vectors of data and an option that indicates whether the function should use `chisq.test()` or `fisher.test()`. Here's an example function:

```
> bivariate.categorical(ii$gender,ii$hs)

    Pearson's Chi-squared test with Yates' continuity correction

data:  table(yy, xx)
X-squared = 131.05, df = 1, p-value < 2.2e-16
> bivariate.categorical(ii$gender,ii$hs,fisher=TRUE)

    Fisher's Exact Test for Count Data

data:  table(yy, xx)
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.4241661 0.5473432
sample estimates:
odds ratio
 0.4819085
```

```
> bivariate.categorical<-function(yy,xx,fisher=FALSE){
+   if(!fisher){
+     tt<-chisq.test(table(yy,xx))
+   } else {
+     tt<-fisher.test(table(yy,xx))
+   }
+   tt
+ }
> bivariate.categorical(ii$gender,ii$hs)

    Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(yy, xx)
X-squared = 131.05, df = 1, p-value < 2.2e-16
> bivariate.categorical(ii$gender,ii$hs,fisher=TRUE)

    Fisher's Exact Test for Count Data

data:  table(yy, xx)
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.4241661 0.5473432
sample estimates:
odds ratio
 0.4819085
```

Use the `presidencies.csv` data file for the following problems.

```
> ii<-read.csv('presidencies.csv')
```

4. Use `tapply()` to find the total duration, in days, that each party (`party`) has held office.

```
> ii$start<-as.Date(ii$start,'%d/%m/%Y')
> ii$end<-as.Date(ii$end,'%d/%m/%Y')
> ii$duration<-as.numeric(ii$end-ii$start)
> tapply(ii$duration,ii$party,sum)
                                Democratic
                                     32099
                    Democratic-Republican
                                      8766
Democratic-Republican/National Republican
                                      1461
                Democratic/National Union
                                      1419
                                Federalist
                                      1460
                               Independent
                                      2865
                                Republican
                                     30680
                Republican/National Union
                                      1503
                                      Whig
                                      2922
```

5. Use `sapply()` to find the object type of each variable.

```
> sapply(1:ncol(ii),function(cc){class(ii[,cc])})
[1] "integer" "factor"  "Date"    "Date"    "factor"  "factor"  "factor"
[8] "numeric"
```

Use the `nhanes.csv` data file for the following problems.

```
> ii<-read.csv('nhanes.csv')
```

6. Use `tapply()` to find the median weight (`bmxwt`) of each age group (`agegroup`), but do this only for the males (`gender`). One way to include the stratification by males is to use `with()` and the `with()` function's `data` argument. For example, using different variables:

```
with(data=subset(ii,hs=='No'),tapply(bpxsy1,bpcat,mean,na.rm=TRUE))
    Elevated Hypertension       Normal
    123.6530     136.7109     108.6307
```

```
> with(data=subset(ii,gender=='Male'),tapply(bmxwt,agegroup,median))
18-34 35-49 50-64 65-69
80.60 86.65 84.20 85.80
```

7. Use `apply()` to find the median weight (`bmxwt`) and height (`bmxht`) of all individuals.

```
> apply(ii[,c('bmxwt','bmxht')],2,median)
bmxwt bmxht
 79.5 166.8
```

8. Use `apply()` to find the median weight (`bmxwt`) and height (`bmxht`) of all females (`gender`). Hint: you can use indexing or `subset()` in combination with the `apply()` function's `X` argument.

```
> apply(ii[ii$gender=='Female',c('bmxwt','bmxht')],2,median)
bmxwt bmxht
 74.0 160.6
```

9. The `tapply()` function can be a useful way to summarize tall data. Use `reshape()` to create a tall version of the NHANES data, for the repeated measurements of diastolic blood pressure. You can follow the example beginning on slide 57 of Module 7 to do this. Order the reshaped data according to `seqn`, and then subset to just the first 24 rows. Now, use `tapply()` to find the mean diastolic blood pressure of each individual. You should obtain 6 values (each value represents the mean diastolic blood pressure of a unique individual).

```
> vv<-c('bpxdi1','bpxdi2','bpxdi3','bpxdi4')
> ss<-ii[,c('seqn','ridageyr',vv)]
> mm<-reshape(data=ss,varying=vv,sep='',direction='long',
+             idvar='seqn',timevar='order')
> mm<-mm[order(mm$seqn),]
> mm<-mm[1:24,]
> tapply(mm$bpxdi,mm$seqn,mean,na.rm=TRUE)
   83732    83733    83735    83736    83741    83742
65.33333 86.00000 70.00000 60.00000 72.66667 70.66667
```