Laboratory Exercise Week 5

Ryan Estes Section 006, 10 pts

09/20/2017

*Directions*:

* Write your R code inside the code chunks after each question.
* Write your answer comments after the # sign.
* To generate the word document output, click the button Knit and wait for the word document to appear.
* RStudio will prompt you (only once) to install the knitr package.
* Submit your completed laboratory exercise using Blackboard's Turnitin feature. Your Turnitin upload link is found on your Blackboard Course shell under the Laboratory folder.

For this exercise, you will need to use the packages mosaic and dplyr.

library(mosaic) # load the package mosaic to use its functions

## Loading required package: dplyr

##   
## Attaching package: 'dplyr'

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

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

## Loading required package: lattice

## Loading required package: ggformula

## Loading required package: ggplot2

##   
## New to ggformula? Try the tutorials:   
## learnr::run\_tutorial("introduction", package = "ggformula")  
## learnr::run\_tutorial("refining", package = "ggformula")

## Loading required package: mosaicData

## Loading required package: Matrix

##   
## The 'mosaic' package masks several functions from core packages in order to add   
## additional features. The original behavior of these functions should not be affected by this.  
##   
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.

##   
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':  
##   
## mean

## The following objects are masked from 'package:dplyr':  
##   
## count, do, tally

## The following objects are masked from 'package:stats':  
##   
## binom.test, cor, cor.test, cov, fivenum, IQR, median,  
## prop.test, quantile, sd, t.test, var

## The following objects are masked from 'package:base':  
##   
## max, mean, min, prod, range, sample, sum

library(dplyr) # load the package dplyr to use data management functions

1. For decades it's been suspected that schizophrenia involves anatomical abnormalities in the hippocampus, an area of the brain involved with memory. The following data bearing on this issue are from Suddath et al. (1990) and were used by Ramsey and Schafer (3rd ed., 2013, p. 31. Display 2.2). The researchers obtained MRI measurements of the volume of the left hippocampus from 15 pairs of identical twins discordant for schizophrenia, i.e, one the twin is affected with schizophrenia The data are displayed in the following table.

### Do not delete this code chunk

# another way to load a small data set using `read.table()`  
schizophrenia <- read.table(header = T, text="  
pair affected unaffected  
1 1.27 1.94  
2 1.63 1.44  
3 1.47 1.56  
4 1.39 1.58  
5 1.93 2.06  
6 1.26 1.66  
7 1.71 1.75  
8 1.67 1.77  
9 1.28 1.78  
10 1.85 1.92  
11 1.02 1.25  
12 1.34 1.93  
13 2.02 2.04  
14 1.59 1.62  
15 1.97 2.08  
")  
is.data.frame(schizophrenia) # check if object `schizophrenia` is a valid data frame

## [1] TRUE

str(schizophrenia)

## 'data.frame': 15 obs. of 3 variables:  
## $ pair : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ affected : num 1.27 1.63 1.47 1.39 1.93 1.26 1.71 1.67 1.28 1.85 ...  
## $ unaffected: num 1.94 1.44 1.56 1.58 2.06 1.66 1.75 1.77 1.78 1.92 ...

schizophrenia

## pair affected unaffected  
## 1 1 1.27 1.94  
## 2 2 1.63 1.44  
## 3 3 1.47 1.56  
## 4 4 1.39 1.58  
## 5 5 1.93 2.06  
## 6 6 1.26 1.66  
## 7 7 1.71 1.75  
## 8 8 1.67 1.77  
## 9 9 1.28 1.78  
## 10 10 1.85 1.92  
## 11 11 1.02 1.25  
## 12 12 1.34 1.93  
## 13 13 2.02 2.04  
## 14 14 1.59 1.62  
## 15 15 1.97 2.08

i. use `mutate()` to create a new variable `diff` which is the difference of the MRI measurements of each pair.   
  
ii. use the pipe `%>%` operator add the new variable `diff` as column to `schizophrenia`.   
  
iii. use `summarise` to compute the average difference of the MRI measurements. Use the pipe `%>%` operator to string multiple functions.   
  
iv. use `summarise` to compute the standard deviation of the difference of the MRI measurements. Use the pipe `%>%` operator to string multiple functions.   
  
v. based on your answers in (iii) and (iv), do you think there is evidence in favor of the initial hypothesis there is difference in the MRI measurements of the volume of the left hippocampus between those affected and unaffected with schizophrenia?

### Code chunk

# start your code  
# i-ii)  
schizophrenia %>%  
 mutate(diff = abs(affected - unaffected))

## pair affected unaffected diff  
## 1 1 1.27 1.94 0.67  
## 2 2 1.63 1.44 0.19  
## 3 3 1.47 1.56 0.09  
## 4 4 1.39 1.58 0.19  
## 5 5 1.93 2.06 0.13  
## 6 6 1.26 1.66 0.40  
## 7 7 1.71 1.75 0.04  
## 8 8 1.67 1.77 0.10  
## 9 9 1.28 1.78 0.50  
## 10 10 1.85 1.92 0.07  
## 11 11 1.02 1.25 0.23  
## 12 12 1.34 1.93 0.59  
## 13 13 2.02 2.04 0.02  
## 14 14 1.59 1.62 0.03  
## 15 15 1.97 2.08 0.11

# iii)  
schizophrenia %>%  
 mutate(diff = abs(affected - unaffected)) %>%  
 summarise(ave.diff = mean(diff))

## ave.diff  
## 1 0.224

# iv)  
schizophrenia %>%  
 mutate(diff = abs(affected - unaffected)) %>%  
 summarise(sd.diff = sd(diff))

## sd.diff  
## 1 0.2128648

# v)  
# Since the Std. Dev. and the Mean are similar, it would suggest that there is little evidence to suggest this hypothesis based on the data.  
  
# last R code line

1. The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey run by the [Centers of Disease Control](http://www.cdc.gov/brfss) in the United States. The BRFSS is designed to identify risk factors in the adult population and report emerging health trends. For example, respondents are asked about their diet and weekly physical activity, their HIV/AIDS status, possible tobacco use, and even their level of healthcare coverage.

We only consider a subset of the original survey that contains 20,000 observations.

### Do not delete this code chunk

cdc <- read.csv("http://www.siue.edu/~jpailde/cdc.csv")  
str(cdc)

## 'data.frame': 20000 obs. of 10 variables:  
## $ ID : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ genhlth : Factor w/ 5 levels "excellent","fair",..: 3 3 3 3 5 5 5 5 3 3 ...  
## $ exerany : int 0 0 1 1 0 1 1 0 0 1 ...  
## $ hlthplan: int 1 1 1 1 1 1 1 1 1 1 ...  
## $ smoke100: int 0 1 1 0 0 0 0 0 1 0 ...  
## $ height : int 70 64 60 66 61 64 71 67 65 70 ...  
## $ weight : int 175 125 105 132 150 114 194 170 150 180 ...  
## $ wtdesire: int 175 115 105 124 130 114 185 160 130 170 ...  
## $ age : int 77 33 49 42 55 55 31 45 27 44 ...  
## $ gender : Factor w/ 2 levels "f","m": 2 1 1 1 1 1 2 2 1 2 ...

i) How many variables are present in this data set? For each variable, identify its data type (e.g. categorical, continuous).   
  
ii) Use `summarise` to compute the average of the variable `weight`.  
  
iii) Use `group\_by` to group the rows by `exerany` (exercise any). Repeat part (ii) on this grouped data. Comment on what you observe in the average weights between groups. Use the pipe `%>%` operator to string multiple functions.

Do not print the data frame (too long, 20K rows), just print the average weights.

iv) Repeat part (iii) but now use the grouping variables `smoke100` and `gender`. Comment on what you observe in the average weights between groups. Use the pipe `%>%` operator to string multiple functions.   
  
v) Obtain a random sample of 1000 rows and save this into `cdc.samp1`.   
  
vi) Repeat parts (ii) to (v) but only using the subset data `cdc.samp1`. Use the pipe `%>%` operator to string multiple functions.   
  
vii) Comment on what differences you observed (if any) between the results in the original sample and the smaller sample?

### Code chunk

# start your code  
# i) There are 10 variables in the data set.  
  
# ii)  
cdc %>%  
 summarise(ave.weight = mean(weight))

## ave.weight  
## 1 169.683

# iii)  
cdc %>%  
 group\_by(exerany) %>%  
 summarise(ave.weight = mean(weight))

## # A tibble: 2 x 2  
## exerany ave.weight  
## <int> <dbl>  
## 1 0 171.5722  
## 2 1 169.0387

# Those who exercise are of less weight. 2 pounds on average to be exact.  
  
# iv-1)  
cdc %>%  
 group\_by(smoke100) %>%  
 summarise(ave.weight = mean(weight))

## # A tibble: 2 x 2  
## smoke100 ave.weight  
## <int> <dbl>  
## 1 0 167.6073  
## 2 1 172.0043

# Those who don't smoke weight less.  
  
# iv-2)  
cdc %>%  
 group\_by(gender) %>%  
 summarise(ave.weight = mean(weight))

## # A tibble: 2 x 2  
## gender ave.weight  
## <fctr> <dbl>  
## 1 f 151.6662  
## 2 m 189.3227

# Females weigh significantly less than males.  
  
# v)  
cdc.samp1 <- sample\_n(cdc, size = 1000)  
  
# vi-ii)  
cdc.samp1 %>%  
 summarise(ave.weight = mean(weight))

## ave.weight  
## 1 170.635

# vi-iii)  
cdc.samp1 %>%  
 group\_by(exerany) %>%  
 summarise(ave.weight = mean(weight))

## # A tibble: 2 x 2  
## exerany ave.weight  
## <int> <dbl>  
## 1 0 173.3019  
## 2 1 169.6735

# Those who exercise are of less weight. 2 pounds on average to be exact.  
  
# vi-iv-1)  
cdc.samp1 %>%  
 group\_by(smoke100) %>%  
 summarise(ave.weight = mean(weight))

## # A tibble: 2 x 2  
## smoke100 ave.weight  
## <int> <dbl>  
## 1 0 168.5291  
## 2 1 173.0385

# Those who don't smoke weight less.  
  
# vi-iv-2)  
cdc.samp1 %>%  
 group\_by(gender) %>%  
 summarise(ave.weight = mean(weight))

## # A tibble: 2 x 2  
## gender ave.weight  
## <fctr> <dbl>  
## 1 f 151.8755  
## 2 m 191.1213

# Females weigh significantly less than males.  
  
# vii) The results were very similar but not consistent.  
  
# last R code line

1. Use sample() to generate rolls from biased coin with .
   1. get a sample of size 10 tosses and tally the results
   2. get a sample of size 30 tosses and tally the results
   3. get a sample of size 100 tosses and tally the results
   4. what do you notice with the proportion of heads in each sample?

### Code chunk

# start your code  
# i)  
coin <- c("Head", "Tail")  
coin.toss10 <- sample(coin,  
 prob = c(0.6, 0.4),  
 size = 10,  
 replace = TRUE)  
tally(coin.toss10)

## X  
## Head Tail   
## 6 4

# ii)  
coin <- c("Head", "Tail")  
coin.toss30 <- sample(coin,  
 prob = c(0.6, 0.4),  
 size = 30,  
 replace = TRUE)  
tally(coin.toss30)

## X  
## Head Tail   
## 19 11

# iii)  
coin <- c("Head", "Tail")  
coin.toss100 <- sample(coin,  
 prob = c(0.6, 0.4),  
 size = 100,  
 replace = TRUE)  
tally(coin.toss100)

## X  
## Head Tail   
## 60 40

# iv)  
# The amount of heads are always well above 50% for each data set. This is because it has a hihger probability of heads.  
  
# last R code line