Midterm Exam: STAT 380, Spring 2017
Penn State

NAME

I have neither given nor received any assistance in the taking of this exam.
Signature:

Instructions:

1. This is a closed book exam.
2. *Turn off all electronic devices*! Any student whose electronic device rings, vibrates, or makes any sound during the exam shall turn in their paper and leave the room immediately.
3. Please verify that your exam paper contains all 6 questions.
4. To earn credit, write clearly, and show your work. When in doubt, explain things clearly. If you need extra paper then please write on the back of your paper.

DO NOT WRITE BELOW THIS LINE

Question	Marks
1	
2	
3	
4	
5	
6	
Total	

Q1 (16 pts) The Behavioral Risk Factor Surveillance Survey: is conducted biannually. We have the data from 1970 and 1990. The following measurements were made on each individual surveyed:

- weight: measured in pounds, with a value of 999 for missing
- height: measured in inches
- age: measured in years
- daysSick: number of days in the past 30 days that health was not good. Values include 1 to 30. None is coded as 88.
- lastCheckup: time since last routine checkup. Values are 1 for within a year, 2 for more than 1 and less than 2 years, 3 for more than 2 and less than 5 years, 4 for more than 5 years, and NA for dont know.
- phone: the respondent was contacted by land line (1) or cell phone (2).

The data are in a data frame called BRFSS. To prepare the data for analysis, write code to perform each of the following operations.

(a) Recode the number of days sick so that 88 is 0.

```
BRFSS$daysSick[BRFSS$daysSick == 88] <- 0
```

(b) Convert a weight of 999 to NA.

```
BRFSS$weight[BRFSS$weight == 999] <- NA
```

(c) Turn lastCheckup into a factor with appropriate labels for the levels.

```
BRFSS$lastCheckup <- factor(BRFSS$lastCheckup,
labels = c('less than 1 yr', '1-2 yrs', '2-5 yrs', 'more than 5 yrs'))
```

(d) Drop all records from the data frame that have a value of NA for the time of the last check up. (Assign the smaller data frame to BRFSS2).

```
BRFSS2 <- BRFSS[!is.na(BRFSS$lastCheckup), ]
```

Q2 (20 pts) Write a function called qcd(), short for quartile coefficient of dispersion. This function takes two arguments: the required x, which is a numeric vector; and the optional na.rm, which indicates whether NAs should be removed from x. The default value for this argument should be FALSE. The function returns the single numeric value that is the ratio of the interquartile range (=75th percentile-25th percentile) to the sum of the lower (25th percentile) and upper (75th) percentiles. Note: the function IQR(x, na.rm=FALSE) returns the interquartile range of the vector x. Also, the function quantile(x, 0.25) returns the 25th percentile. In addition, check that the input for x is numeric and if not terminate execution and provide an informative error message.

```
qcd <- function(x, na.rm = FALSE) {
    if(!is.numeric(x)) stop(‘x must be a number’)
    if(na.rm) x <- x[!is.na(x)]
    iqr <- IQR(x, na.rm = na.rm)
    q25 <- quantile(x, 0.25)
    q75 <- quantile(x, 0.75)
    return(iqr/(q25 + q75))
}
```

Q3 (16 pts) What is the value printed to the console for each of the following expressions?

(a)
```
> x = c(-1, -1, -1, -1, 1, -1)
> y = cumsum(x)
> y
```

```
[1] -1 -2 -3 -4 -3 -4
```

(b)
```
> z = which(y < -10)
> z
```

```
integer(0)
```

(c)
```
> z[length(z)]
```

```
integer(0)
```

(d)
```
> x[z]
```

```
numeric(0)
```

Q4 (12 pts) Write a function that uses Monte Carlo to approximate the distribution of the sample proportion of Heads from n coin tosses of a coin (which may or may not be a fair coin). The function should take as input the sample size n, the probability of heads p, the number of replicates (repetitions of the experiment) B and a boolean `plotHist` that if TRUE prints a histogram of the samples and if not, does not. If the user does not input a value for B, default to the one we often use in class, that is, set $B = 10,000$. The output of the function should be the mean and standard deviation of the distribution.

```
mc <- function(n, p, B = 10000, plotHist = TRUE) {
    numHeads <- rep(NA, B)
    for(i in 1:B) {
        flips <- sample(0:1, size = n, replace = TRUE, prob = c(1-p, p))
        numHeads[i] <- sum(flips)
    }
    if(plotHist) hist(numHeads)
    return(list(mean = mean(numHeads), sd = sd(numHeads)))
}
```

or

```
mc <- function(n, p, B = 10000, plotHist = TRUE) {
    numHeads <- rbinom(B, n, p)
    if(plotHist) hist(numHeads)
    return(list(mean = mean(numHeads), sd = sd(numHeads)))
}
```

Q5 (20 pts)

(a) Suppose you want to analyze daily temperatures at 3 different locations and that the number of days where data are available at these locations is quite different. What kind of data structure would you use to store all these data and why?

A list. Data frames require every vector to be the same length, lists do not.

(b) Suppose you have a dataset of heights (in) and weights (lbs) of adults. You know there are 25,000 rows in the data. When you make a scatterplot, this is what you see:

`plot.png`

i. Why are there fewer than 25,000 points on the plot? What can you do to make more points visible?

The data are integer valued, so many points lie on top of each other. We can “jitter” the points, adding a small amount of noise in all directions so that the points appear more spread out.

ii. How can you improve this plot to make it more informative for a reader?
Axis labels, specify units, plot title, ...

Q6 (16pts) Consider the following list, aList:

```
aList
  $x
  [1] "a" "b" "c" "d" "e"
  $mat
       [,1] [,2]
  [1,]    8    5
  [2,]    7    4
  [3,]    6    3
$zz
$zz$x
[1] 1 2 3
$zz$y
[1] 7 8
91011
$zz$z
[1]  TRUE  TRUE FALSE FALSE  TRUE
$one [1] 100
```

Write down what will appear at the console when R evaluates each of the following expressions (note that some expressions may result in an error message):

(a) `length(aList)`

```
[1] 4
```

(b) `aList$mat + aList$one`

```
     [,1] [,2]
[1,]  108  105
[2,]  107  104
[3,]  106  103
```

(c) `length(aList[["zz"]][1])`

```
[1] 1
```

(d) `sapply(aList$zz, min)`

```
x y z
1 7 0
```