R Exercises

MATH 4939

December 31, 2018

1. What output will the following R script produce? Explain briefly why.

   ```
   x <- c(TRUE, FALSE, 0L)
   typeof(x)
   ```

2. What output will the following R script produce? Explain briefly why.

   ```
   TRUE | NA
   ```

3. Let x be defined as:

   ```
   x <- c('0','10','5','20','15','10','0','5')
   ```

   Write an R function that would turn `x`, or other vectors similar to it, into a factor whose ordering corresponds to the numerical ordering of `x`.

4. In R, let `x <- 1:5`. What output would `x[NA]` produce? What output would `x[NA_real_]` produce? Describe the reason for the difference, if any.

5. In R, describe the result of subsetting a vector with positive integers, with negative integers, with a logical vector, or with a character vector?

6. In R, what's the difference between [, [[, and $ when applied to a list?

7. In R, when subsetting with [, when should you use drop = FALSE? Include arrays and factors in your discussion.

8. In R, If `x` is a matrix, what does `x[] <- 0` do? How is it different from `x <- 0`? Why?

9. In R, how can you use a named vector to relabel a categorical variable?

10. In R, if mtcars is a data frame, why does mtcars[1:20] return an error? How does it differ from the similar mtcars[1:20,]?

11. In R, if df is a data frame, what does df[is.na(df)] <- 0 do? How does it work?

12. Create the vector (20,19, ... ,2,1) in R.

13. Create the vector (1,2,3, ... ,19,20,19,18, ... ,2,1) in R.

14. Create the vector (4,4, ... ,4,6,6, ... ,6,3,3, ... ,3) in R, where there are 10 occurrences of 4, 20 of 6 and 30 of 3.

15. Write an expression in R to calculate the following $\sum_{i=10}^{100}(i^3 + 4i^2)$

16. Generate in R a vector of 30 labels: 'label 1', 'label 2', ... 'label 30'

17. Let `y <- sample(1000, 30, replace = TRUE)`. Write an expression in R to determine how many elements of `y` are divisible by 2.

18. Let `y <- sample(1000, 30, replace = TRUE)`. Write an expression in R to determine how many elements of `y` are within 200 of the maximum value.

19. Let `y <- sample(1000, 30, replace = TRUE)`. Write an expression in R to determine how many elements of `y` are less than the previous element.

20. Let `y <- sample(1000, 30, replace = TRUE)`. Write an expression in R to determine how many elements of `y` are an exact square.

21. Suppose data for a variable in R representing dollars has been entered in a variety of formats: '$1,000.00','1000.00','$1'. Write a function in R that transforms the variable to a numeric variable in dollars to the nearest cent.

22. Write a function in R that takes a character string and collapses multiple adjoining blanks to a single blank.

23. Use the site Gapminder.org to download at least three longitudinal variables into separate data sets. Merge the data sets into one for which each row represents one country and year and contains the values of each of the three variables you downloaded.

24. Write a function in R that removes from a data frame every variable whose name starts with the letter 'X' and ends in a number.

25. Write R code to create a 6 by 10 matrix of random integers in R as follows:

```
set.seed(75)
m <- matrix(sample(10, 60, replace = T), nrow = 6)
```

Write a function to find the number of entries in each row that are greater than 4. Use the `apply` function.

26. Let `mat` be a matrix of integers in R. Write a function to find how many rows have exactly two instances of the number 7. Use the `apply` function.

27. Describe the difference in R between `paste(x, y, sep = ':')` and `paste(x, y, collapse = ':')`

28. Using the `hs` data set in the spida2 package, create a plot with two panels showing histograms displaying the distribution of school sizes in the Public and in the Catholic sectors. Use the functions `capply` and `up` in the spida2 package. You may also use any other approach to compare with the use of `capply` and `up`.

29. Using the `hs` data set in the spida2 package, create a plot with two panels showing histograms displaying the distribution of sample sizes in each school in the Public and in the Catholic sectors. Use the functions `capply` and `up` in the spida2 package. You may also use any other approach to compare with the use of `capply` and `up`.

30. Using the `hs` data set in the spida2 package, create a plot with two panels showing scatterplots displaying the relationship between mean mathach and mean ses in each school in the Public and in the Catholic sectors. Explore reasonable transformations and regression lines: linear and non-parametric in the plots. Use the functions `capply` and `up` in the spida2 package. You may also use any other approach to compare with the use of `capply` and `up`.

31. Describe the difference in R between a 'generic function' and a method.

32. Which of the following R expressions result in the following output?

```
[1]  8
```

(Write 'Y' for yes, 'N' for no, and 'D' or blank for 'do not know'. +1 for a correct answer, -1 for a wrong answer and 0 for 'D')

_____ `"+"(5,3)`

_____ `"/"(16,2)`

_____ `2^3`

_____ `4 + 4`

_____ `"^"(3,2)`

33. Suppose we run this command:

```
a <- matrix(1:8, rnow = 2)
```

Then which of the following R expressions result in the following output?

```
[1]  8
```

(Write 'Y' for yes, 'N' for no, and 'D' or blank for 'do not know'. +1 for a correct answer, -1 for a wrong answer and 0 for 'D')

_____ `a(8)`

_____ `a[8]`

_____ `a(2,4)`

_____ `a[4,2]`

_____ `a(2,4)`

34. Suppose you wish to estimate the relationship between income, Y, and education, X. Because of heteroscedasticity and curvature in the relationship you choose to fit a linear model using the log of Y:

```
fit <- lm( log(Y) ~ X, data)
```

Write the R code you would use to plot the estimated increase in income associated with an extra year of education as a function of years of education. It is not necessary to include error bars in the plot.

35. What output with the following script in R produce. Explain why.

$$'1' == \texttt{c}(\texttt{TRUE}, 1)$$

36. Install the microbenmark package in R and use it to compare the time use using a for loop versus vectorization to assign `y <- sin(x)` for a long vector `x`. Discuss your results.

37. Write a function in R that finds the position of the maximum value in a vector using a for loop. Write at least two other functions that do the same thing using different approaches to the problem. Use the microbenchmark package to compare their speeds.

38. Write a function in R that reports whether two sets are disjoint, identical, nested with the first a subset of the second or vice-versa, or otherwise.

39. Write a function in R that generates n points around a circle of radius r centered at the origin.

40. Write a function in R that generates n points around the concentration ellipse ...

41. Write a function in R that generates points on the isoquant of the multivariate T ...

42. Write a tutorial that illustrates every form of indexing in R. Include logical indexing of data frames and include indexing in replacement functions.

43. Compute the truth table for logical OR and logical AND. Note that R has 3 logical values: TRUE, FALSE, as.logical(NA). Write a clear explanation of the logic behind the truth tables.

44. What is the difference between the expressions `x == TRUE` and `isTRUE(x)`. Comment.

45. Write an R function that returns the prime factors of an integer in a vector in which repeated prime factors are repeated according to their multiplicity as prime factors. Try to make your function efficient by avoiding unnecessary operations.

46. Write an R function that takes a vector of character strings, factors or numeric values representing an amount of money that might be entered in a variety of formats for Canadian currency, e.g. “1,000.21”, “$20”, “C23.01”, ”22CDN”, and return a numeric value.

47. Write an R function that takes a street address, e.g. “4700 Keele Street, North Ross 520” or “66-2 Bloor St. W”, and returns a character vector with the street address, e.g. “4700” or “2”, the street name, the unit type (using “Unit” as a default) and the unit name or number. Demonstrate how your function works on a variety of addresses.

48. Consider classlists for four sections of a second year statistics course STA200 (at URL) and two classlists for a third year statistics course STA300. Without any direct editing of the classlists do the following: 1. Write a function that transforms each input classlist into a data frame with useful variables on the program of each students. Note that information on program is encoded in a single column that can contain information on a number of distinct variables. You need to use string manipulation functions, e.g. sub, gsub, strsplit, to turn this column into useful variables. Note that a space is usually a delimiter between subfields but sometimes not. You might need to preprocess the strings before splitting them into subfieds. 2. Is there evidence that a different proportion of students in the 2nd year course go on to study statistics in the 3rd year course? 3. Is there evidence that this remains true when adjusting for the program of students in the 2nd year course? 4. Are there conversions, i.e. students who change their majors to statistics? Do they come disproportionately from some sections instead of others?

49. Something on SVD including plotting axes. Every non-singular linear transformation maps a unit sphere to an ellipsoid that can be taken to be the unit sphere of an ‘elliptical metric’. In two dimensions choose an appropriate transformation, A, and use a suitable graphic to illustrate the geometric interpretation of the components of the SVD: A = UDV’, that is, the column vectors of U and V and the diagonal elements of D.