BMI 713: Computational Statistics for Biomedical Sciences

Assignment 1

September 9, 2010 (due Sept 16 for Part 1; Sept 23 for Part 2 and 3)

1 Basic

1. Use the `:` operator to create the vector $(1, 2, 3, 4, 5)$. Store this vector in a variable.
2. Extend the vector you just created with the values $(6, 7, 8)$. Use the `c` function.
3. Shrink the vector you just created by removing the first element. One could also use the `[]` operators with a negative index to remove an element.
4. Set the first value in the remaining vector to 9.
5. Decrease the value of every element in the vector by 1. Do this using a single arithmetic operation.
6. Create a vector v of length 20 such that all the elements of v are 1. Hint: use the function `rep()` (Can you do the same thing without using the function `rep()`?).
7. Try `rep(1:4,each=2)` and `rep(1:4,times=2)`. Are the results the same? If not, what is the difference?
8. Use R to calculate $\sqrt{2}$, 2^{10}, $\log(10)$, e^2, $\cos(\pi)$, $\sin(\pi)$ (Note: π in R is called `pi`).
9. Given a vector `v = (1,2,3,4,5)`. What happens if you try `1/v`? How about `exp(v)`? How about `v*v)`?
10. Create a vector `v=(2,2,5,7,9,3,4)`. Try `v[v==2]`, `v[v>2]`, `v[v=<5]`. What are the results?
11. Create a vector `v = (2,3,6,11,18)`. What happens if you try `diff(v)` and `diff(v,lag=2)`?
12. Create a vector `v = (15.9,21.4,19.9,21.9,20.0,16.5,17.9,17.5)`. What is the smallest value in `v`? What is the largest value? How many are greater than 18? Make a new vector with only those bigger than 18.
13. What does R return if you try `1/0`? How about `1/0+1/0`, `0/0`, `1/0-1/0`? (Note: `Inf` means infinity and `NaN` means "not a number".)

2 Intermediate

1. Write a program to print out the numbers 1,...,10, separated by tab or a new line. Compute the sum of squares of those numbers.
2. Try `v = rnorm(20)`. What is the length of the vector `v`? What is the summation of `v`? (Hint: Use the function `sum`.) What is the mean? (Hint: Use the function `mean`.) What is the variance and standard deviation? (Use the function `var` and `sd`.)
3. Try `v = rnorm(20,mean=100)`. What is the mean of `v`? What is the variance? Compare this mean with the mean of the previous problem. Are these two means similar? How about the variances? Are they similar? Try `v = rnorm(20,mean=100,sd = 3)` and compare with previous results. (Note: `rnorm()` can generate random numbers from the normal distribution.)

4. Suppose that you track your commute times for two weeks (ten days), recording the following times (in minutes)

   `17 16 20 24 22 15 21 15 17 22.`

   Find the maximum, minimum, mean and standard deviation of your commute times. How many days was the commute time less than 20 minutes? How many days was the commute time between 17 and 21 minutes (including 17 and 21)?

5. Let `v` be the vector `v <- c(9, 2, 8, 8, 5)`.

   (a) Sort `v` using the `[]` operators. (The `order` function will be useful here.)

   (b) Select a random subsample of `v`. (The `sample` function will be useful here.)

   (c) Use the `[]` operators and a boolean operation to select the elements in `v` which are greater than 5.

6. Consider the matrix `M <- matrix(1:9, ncol=3)`.

   (a) Use the `[]` operators to select only the first row of `M`.

   (b) Use the `[]` operators to select the second and third columns of `M`.

   (c) In the previous section, you used matrix multiplication to compute some rows and columns of a matrix. Now you have used the `[]` operators to *select* rows and columns in a matrix. What is the difference? (*Hint*: what if you wanted to alter values in a matrix?)

   (d) Use the `[]` operators to select 3 copies of the first row of `M`. *I.e.*, create the matrix

   $$\begin{pmatrix} 1 & 4 & 7 \\ 1 & 4 & 7 \\ 1 & 4 & 7 \end{pmatrix}$$

   using only the `[]` operators.

3 Advanced

1. `x = c(2, 10, NA, 5, NA); y = c(NA, 5, 4, 3, NA)` Write a program to get a list of positions where no values are missing. `is.na()` is a useful command here.

2. `x = matrix(c(5,3,2,9,3,14,7,6,8),3,3)` What a program to find the location of the maximum value of the matrix.

3. Create and sort the matrix

   $$M = \begin{pmatrix} 2 & 5 & 8 \\ 3 & 6 & 9 \\ 1 & 4 & 7 \end{pmatrix}$$

   by its first column.

4. More matrices

   (a) Create a 4x4 zero matrix using the `matrix` function.

   (b) Create a 4x4 identity matrix using the `diag` function.

   (c) Add the two previous matrices.

   (d) Use the `lower.tri` function to create a lower triangular matrix (all entries below the diagonal are 1). (Note: the `lower.tri` function creates a matrix of boolean (`TRUE` and `FALSE`) values. Use this boolean matrix in the `[]` operators to select the lower triangular portion of another matrix you create.)

   (e) Make a triangular matrix where the entries along the diagonal are also 1.

(f) Compute the following matrix product:

$$\begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

The standard multiplication operator (`*`) won't work here; you must use the special matrix multiplication operator `%*%`. What product would give the second column? The first row?

5. Assume that you know that the mean retail price of certain product nationwide is \$17. Suppose that you want to study the retail price of the product in Boston. You randomly choose 8 stores in Boston and record their retail prices of the product as `v = (15.9,21.4,19.9,21.9,20.0,16.5,17.9,17.5)` (in dollars). Use the function `t.text()` to calculate the P-value of the following alternative hypotheses (the null hypothesis is that the mean retail price in Boston is the same as that nationwide).

   (a) The mean retail price in Boston is not equal to that nationwide.

   (b) The mean retail price in Boston is less than that nationwide.

   (c) The mean retail price in Boston is more than that nationwide.

   Note: use the `t.text()` with proper `mu` value `alternative` value.

6. Comparison of different normal distributions.

   (a) Generate 100 random numbers from the standard normal distribution. Save these numbers in a vector `v1`.

   (b) Generate 100 random numbers from the normal distribution with mean 3 and standard deviation 1. Save these numbers in a vector `v2`.

   (c) Generate 100 random numbers from the normal distribution with mean 3 and standard deviation 5. Save these numbers in a vector `v3`.

   (d) Plot and compare the histograms of the three vectors `v1`, `v2` and `v3`. Does the shapes of hisgograms look similar? What looks different?

   (e) Plot and compare the boxplots of the three vectors `v1`, `v2` and `v3`.

   (f) Test whether the means of `v1` and `v2` are equal using t test. What are the null hypothesis and the alternative hypothesis (Optional question)?

7. Suppose you observed an event at the following times: $t = (1, 4, 5, 12, 15, 17, 33)$. You are interested in studying the time between events. Convert the vector t of event times into the vector d of time differences using a single operation. (*I.e.*, create a vector d such that $d_i = t_{i+1} - t_i$. d will be one element shorter than t.)

8. Let v be a vector of length n. `R` contains a function named `cumsum` which will compute the vector of partial sums of `v`. *E.g.*, if $v = (1, 2, 3, 4, 5)$ then the vector of partial sums is

$$p = (1, 1+2, 1+2+3, 1+2+3+4, 1+2+3+4+5) = (1, 3, 6, 10, 15).$$

   Without using `cumsum`, produce the same vector of partial sums using a single arithmetic operation. (*Hint*: Construct a suitable matrix, then use matrix multiplication to produce the vector of partial sums.)

9. Consider the "rotated" vector $v = (5, 6, 7, 8, 9, 1, 2, 3, 4)$. As you can see, every element in the vector is in order—$v_i < v_{i+1}$—for all i except one. Using a single boolean operation, determine the position i such that $v_i \not< v_{i+1}$.