R Exercises

Name:

Student Number:

**Exercises 1: Operators**

1. Create a new object called “x” and assign it a value 13.4. Write your command line.

Answer:

1. Divide x by 2 and assign the result to a variable called “y”. What is the value of y?

Answer:

1. Perform log10-transformation on “x” and add y. Then put the value to “z”. What is the value of z?

Answer:

1. The variable z has many digit numbers. Wrote the command line to round the z value to the third digit and print a string like: z value = ???. What is your command?

Answer:

1. Write a logic operator to see whether the value of z is at the range of 2 and 10?

Answer:

**Exercise 2: vectors and factors**

One study aims to characterize the number of lean/obese individuals in the cohort. They measured height and body weight for every individual and you can load the data by dragging it into R or use the command

> setwd(‘your working direction”)

> load('height.Rdata')

> load(‘weight.Rdata’)

Two variables called “height” and “weight” are loaded into R.

1. How many individuals they measured in their study?

Answer:

1. Which function you use to get the sample size?

Answer:

1. Assuming the order of samples is the same between height and weight. Please calculate the body max index for each individual, using the formula: bmi=weight/height2 and assign it into the variable bmi. Please fill the summary statistics of BMI in the following table:

|  |  |  |
| --- | --- | --- |
|  | Write down your command line | Output value |
| Minimum value |  |  |
| Maximum value |  |  |
| Average value |  |  |
| Median value |  |  |
| Standard deviation |  |  |
| How many individuals are obese(bmi≥30) |  |  |
| How many individiauls are overweight (25≤bmi<30) |  |  |

9. Generate a **factor** called “group” and classify each individuals to the group value of “underweight”, “normal”, “overweight” and “obese” based on their bmi. (3 points)

* If your BMI is less than 18.5, it falls within the underweight range.
* If your BMI is 18.5 to <25, it falls within the normal.
* If your BMI is 25.0 to <30, it falls within the overweight range.
* If your BMI is 30.0 or higher, it falls within the obese range.

Hit: you can first create a character vector “group” with the same length. Find out which individuals (index) are underweight and assign the value “underweight” to those individuals and so on …..; To change group to a factor and order the levels.

Alternatively, you can use function cut().

Copy& paste your command line below:

**Exercise 3: Matrix and data frame**

10. Use the variable height, weight and BMI and make a matrix called “data” that combines the height, weight and BMI data for each individual. Also add the sample ID to the rows, like “sample1”, ‘sample2”, “sample 3”, …..

Copy and paste your codes below

11. What are the column names of data?

Answer:

12. Please find out the values for the sample54?

Answer:

Height:

Weight:

BMI:

13. Please find out the height value of the individual with the highest BMI?

Answer:

14. Load the gender information from newdata.Rdata. What is the data structure of gender?

Answer:

15. Merge two datasets (newdata and data) into one data frame (mydata). Please pay attention to the orders of samples in both datasets! Check their row names! Are they in the same order? Please find out the average BMI value for female.

Answer:

Please save mydata!! You can use the command below

*save(mydata, file='…direction…/mydata.Rdata').*

**Exercise 4: Association analysis**

16. Transcriptome analysis has approved to be a powerful tool to identify candidate genes for diseases. Now you will perform an association analysis to investigate the association between liver gene expression and the degree of non-alcoholic steatohepatitis (NASH) in 60 obese individuals. Firstly, you need to load the expression data from the file (expression.txt) and convert the data into a matrix: You should generate a 1000x60 matrix, with gene at row and sample at column. Please note that gene names read like ENSG00000083857. This stands for unique ensembl gene ID (www.ensembl.org). You also need to load phenotype data in the same way. Now you need to perform non-parametric correlation analysis (spearman correlation) between the 1st gene and the grade of NASH (NASH\_grade), **after regress out the effect of age and gender**. Hit: (1) firstly retrieve the expression of the 1st gene (2) then use linear model to regress out the effect of age and sex and retrieve the residuals; (3) compute the spearman correlation between the residuals and NASH\_grade

Answer:

Spearman correlation coefficient (r) =

P value =

**Exercise 5:**

Write a loop to repeat the above analysis for all 1,000 genes and save results to a data.frame with three columns: geneID, r.value and p.value.

17) How many genes significantly associated with NASH grade at P<5x10-5

Answer:

18). Which gene shows the strongest association to NASH grade.

Answer:

Gene Name =

R value =

P value =

**Exercises 6:**

1. In the exercise 5, you have identified transcripts that are associated to NASH\_grade. Can you make a graph to show the most significantly associated gene with NASH grade. You can consider which type of plot you can use. Copy your plot below.
2. Make a heatmap for top 50 genes that are significantly associated with NASH grade based on the results of Exercise 5. Copy your plot below