M.Sc. in Data Science

Course: Probability and Statistics for Data Analysis

Instructor: Ioannis Vrontos (vrontos@aueb.gr)

Grader: Constandina Koki (kokiconst@aueb.gr)

Assignment 1 Deadline: 20 November 2017

Exercise 1

Consider the following events in the toss of a single die: A: Observe an odd number, B: Observe an even number, C: Observe 1 or 2.

(a) Are A and B independent events?

(b) Are A and C independent events?

Exercise 2

Males and females are observed to react differently to a given set of circumstances. It has been observed that 65% of the females react positively to these circumstances, whereas only 45% of males react positively. A group of 50 people, 35 female and 15 male, was subjected to these circumstances, and the subjects were asked to describe their reactions on a written questionnaire. A response picked at random from the 50 people was negative. What is the probability that it was that of a male?

Exercise 3

An industry is planning to produce 2 new products, A and B. The probability that product A will be successful given that at the same time a competitor will produce an item similar to A, is 0.4, whereas the probability that the product A will be successful given that no other company will produce similar to A item is 0.7. The probability that a competitor will present a similar to A product is 0.3. The probability that the product B is going to be successful given that the product A is successful is 0.6. Finally the probability that product B is going to be successful is 0.4.

(a) What is the probability that product A is going to be successful *and* a similar to A product is going to be presented by the competitor?

(b) What is the probability that product A is going to be successful?

(c) What is the probability that a similar product to A is going to be presented by the competitor, given that the product A is successful?

Exercise 4

The daily weight of biological waste (in tones) that a biological purification facility uses is a continuous random variable X that has a probability density function given by

$$f(x) = \begin{cases} c(4-2x) & if \;\; 1 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) What is the value of c?

(b) Find the cumulative distribution function of X.

(c) What is the probability that the weight of waste is 1.5 tones at most?

(d) Compute the mean value of X.

Exercise 5

If X is an exponential random variable with parameter λ = 1 compute the probability density function of the random variable Y defined by Y = log X. [Hint: Work with the cumulative distribution function(cdf) and then compute the probability density function (pdf)]

Exercise 6

1. A randomly chosen IQ test taker obtains a score that is a normal random variable with mean 100 and standard deviation 15. What is the probability that the score of such a person is (a) above 125 (b) between 90 and 110 (c) submit the R commands that was used to answer questions 1(a) and 1(b).
2. A multiple-choice examination has 15 questions, each with five possible answers, only one of which is correct. Suppose that one of the students who takes the examination answers each of the questions with an independent random guess. (a) What is the probability that he/she answers at least three questions correctly? (b) submit the R commands that was used to answer question 2(a).

Exercise 7

Use R in this exercise and submit your R code that was used to answer the questions.

From the list of distribution studied select one asymmetric discrete, call it X, and one asymmetric continuous, call it Y. For each of the two distributions generate two random samples of size 100 and 10000. Each time you generate a random sample use the command set.seed(.) using the same unique number of your choice (so that your results are reproducible). For all four generated data sets:

(a) Provide a graphical representation (visualization) of your data.

(b) Get an estimate of the: mean, sd, median, IQR. What are the theoretic values of these statistics when the true distribution is used?

(c) What is the proportion of the sample data in the region:

$$[\min(\text{mean}, \text{median}),\ \max(\text{mean}, \text{median})]$$

What is the theoretic proportion of this region when the true distribution is used?

(d) What is the quantile of the sample data that are at or below the lowest 1%? What is the respective theoretic quantile when the true distribution is used?

(e) What is the quantile of the sample data that are at or above the upper 1%? What is the respective theoretic quantile when the true distribution is used?

(f) Comment on the discrepancies between theoretic and sample evaluated statistics in the previous questions (b)-(e).