Statistical Methods Fall 2021

Assignment 2: Probability, Normality, CLT and Law of Large Numbers

Deadline: see Canvas

Topics of this assignment
The exercises below concern topics that were covered in Lectures 2, 3 and 4: probability, including the Law of Total Probability and Bayes' Theorem, random variables, model distributions, the Central Limit Theorem (CLT), and the Law of Large Numbers (see the respective sections in Chapters 3, 4 and 5 of the book and the handouts of Lectures 2, 3, 4). Before making the assignment, study these topics.

How to solve and present the solutions to the exercises?
See the first page of Assignment 1 and the example file about the required format.

Theoretical exercises

Exercise 2.1 Take a look at Exercise 1.3 from the set of additional exercises (link) and consider the same setting here.

a) What is the probability that a random person, which does the test, gets a positive result? Explain the difference between this probability and the probability you are asked to calculate in the Exercise 1.3 from the set of additional exercises.

b) Solve Exercise 1.3 from the additional exercises (link).

c) Are the two events that *a person has cancer* and that *the test is positive* dependent? Does the fact that a test result was positive increase the risk of having cancer, when compared to the probability that a random individual from the population has cancer?

Hand in: Answers with your calculations, explanations, and motivations.

Also, mention which formulas you are using and state the formulas you are using.

Warning: it is possible that you receive 0 points for certain parts of the exercise if you don't give a motivation/explanation!

Exercise 2.2 Suppose you walk to a bus stop to catch a bus; busses of the line you take arrive there every 15 minutes. You don't look at your watch and arrive at the bus stop totally at random. For simplicity, we assume that all arriving times (you/the busses) are rounded down to whole minutes. When you and the bus arrive simultaneously, you will be able to catch it.

a) Describe the probability space that models the above-described waiting time experiment, i.e. sample space and the related probabilities.

Suppose now that you still start walking towards the bus stop at a completely random time; but that you *do* look at your watch at the beginning of your journey. Knowing your walking pace and the distance to the bus stop, you know by how many minutes you will miss the previous bus. If you would miss the bus by 4 or less minutes (i.e. hypothetically, the waiting time would be 11 minutes or more), you decide to hurry up; in this case, you will still be able to catch it. Otherwise, if you would miss the bus by 5 or more minutes (i.e. waiting times 10 minutes or less), you simply go on with your normal walking pace.

b) Let the random variable X model your waiting time at the bus station in the just-described situation. Calculate the probability that you will need to wait for at least 5 minutes.

c) Calculate the expectation of X.

d) Calculate the variance of X.

e) Suppose you walk to the bus stop in the previously described way year 160 times per year. Denote by $X_1, \ldots, X_{160}$ your waiting times on all these days; assume they are independent of each other. Describe the (approximate) distribution of your average waiting time $\bar{X}_{160} = \frac{1}{n}\sum_{i=1}^{160} X_i$ across the whole year.

Note: only if you are not able to derive the expectation in c) you may continue in d) and e) with the ***wrong*** *value $E(X) = 5$, and if you are not able to derive the variance in d), you may continue in e) with the* ***wrong*** *value $var(X) = 4$.*

Make formal calculations in all parts and do not apply mere reasoning!

Hand in: Answers with your calculations and explanations.

R-exercises *Hints concerning R:*

- Recall that a simple random sample of size n from a set of values `x` can be drawn in R using the function `sample(x,n)`. By default, the sample is drawn without replacement; by setting the additional parameter `replace` to `TRUE`, the sample is drawn with replacement. This function can be used to simulate a die.
- A sample from a certain distribution can be obtained in R with the function `rdist(n,par)` where `dist` stands for the name of the distribution, `n` for the sample size, and `par` for the relevant parameters: `x=rnorm(50,5,1)`, `x=rexp(25,1)`, `x=runif(30,-1,1)`, `x=rt(10,df=5)`, `x=rchisq(25,df=8)`. For example, the function
  `rnorm(n,mean,sd)` generates a sample of size `n` from the normal distribution with expectation `mean` and standard deviation `sd`. The parameters of the other distributions are documented in the help-function.
- A normal QQ plot can be obtained with `qqnorm(x)`, histograms with `hist(x)`, and boxplots with `boxplot(x)`.
- The command `dnorm(u)` computes the value of the probability density function of the standard normal distribution in `u`. For non-standard normal distributions adjust the arguments of the function.
- The command `lines(x,y)` joins the corresponding points in the vectors `x` and `y` with line segments. This is useful to draw a curve on top of an existing plot. Similarly, `abline(a,b)` draws the line $a + bx$ on top of an existing plot. Otherwise specify `type="l"` in the parameters of the function `plot()`.
- To concatenate text and numbers (useful for titles of plots) use the R-function `paste()`.
- Use the command `set.seed(...)` to make your results based on the generated samples reproducible.

Exercise 2.3

a) Generate the following samples and make for each of them a normal QQ plot:

   (i) one sample of size 115 from the chisquared distribution with degrees of freedom 2;

   (ii) one sample of size 105 from the t-distribution with 4 degrees of freedom.

   Evaluate the usefulness of the normal distribution as a model distribution for both samples based on the QQ plots. Comment briefly on each plot and each peculiarity.

   Hint: in a normal QQ plot, data are compared to a theoretical normal distribution.

   Hand in: the 2 plots concisely presented using the command `par(mfrow=c(1,2))`, and your answers.

b) Generate the following samples and make for each of them a histogram and a boxplot.

   (i) one sample of size 115 from the chisquared distribution with degrees of freedom 2;

   (ii) one sample of size 105 from the t-distribution with 4 degrees of freedom.

   Relate the peculiarities visible in the histograms to what you see in the corresponding boxplots, and describe your findings. In particular, address the heaviness of the tails, symmetry, and outliers.

   Hand in: the 4 plots concisely presented using the command `par(mfrow=c(2,2))`, and your answers.

c) Answer for each of the data sets below the following question: "Is it reasonable to assume that the data come from a normal distribution?" In each case choose from the two answers: "Obviously not from a normal distribution" or "Normality cannot be excluded". Base your answer on histograms, boxplots and normal QQ-plots.
Also, for each dataset, point out the peculiarities of each sample by comparing the histogram, boxplot, and QQ-plot with each other. Indicate whether you detect (some/all) peculiarities in some/all of these diagnostic plots.

(i) `titanic3.csv`: Data about many passengers of the Titanic; we are going to analyse the passengers' ages.
More information on http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic.html
Hints: use `titanic <- read.csv("titanic3.csv")` *to read the dataset.*
Then use `titanic$age` *to obtain the passengers' ages.*

(ii) `diabetes.csv`: Data from a study about understanding the prevalence of obesity, diabetes, and other cardiovascular risk factors; we are going to analyse the individuals' total cholesterol values.
More information on http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/diabetes.html
Hints: use `diabetes <- read.csv("diabetes.csv")` *to read the dataset.*
Then use `diabetes$chol` *to obtain the individuals' total cholesterol values.*

(iii) `vlbw.csv`: Data about newborn babies with very low birth weight; we are going to analyse the babies' birth weights.
More information on http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/vlbw.html
Hints: use `vlbw <- read.csv("vlbw.csv")` *to read the dataset.*
Then use `vlbw$bwt` *to obtain the babies' birth weights.*

Hand in: Present for each data set: a suitable histogram, boxplot and QQ-plot, your answer to the question, and a short motivation of this answer. Also, the explanation on whether you identified some/all peculiarities in some/all plots for each dataset. Use the function `par(mfrow=c(1,3))` *to print the three plots next to each other. Adjust the size of the figure so that the ratio becomes approximately 1:3, and each plot is more or less square.*

Exercise 2.4 Study the *R*-function `diffdice` from the file `function02.txt`. Load it by using the command `source("function02.txt")`.

a) Consider two dice and the random variable 'the absolute difference of two die rolls'. Illustrate the Law of Large Numbers for this random variable by considering 'the mean of the absolute difference of two die rolls' in n trials for different values of n and making a plot similar to the one on slide 24 of the Lecture 3 handout.
Hint: n trials means that both of two dice are rolled n times, and for each of the trials the absolut difference is calculated.
You may use (without proof) that the theoretical expected value of the absolute difference is about 1.9444.

b) Use the function `diffdice` to find an approximate value of expectation of the random variable 'the absolute difference of two die rolls' and the probability of the event 'the absolute difference of two die rolls is 3'.

c) Use the function `diffdice` to graphically illustrate the Central Limit Theorem for the random variable 'the mean absolute difference of two dice rolls after n trials' , by making 4 plots similar to the 4 plots on slide 15 of the Lecture 4 handout.
You may use (without proof) that the standard deviation of the ***absolute difference of two*** *dice rolls is approximately* 1.4326.

d) Explain briefly why the 4 plots of part c) illustrate the Central Limit Theorem in the present context.

Hand in: Properly described plots (part a and c), answers with motivation (parts b and d).