BMI713 Problem Set 1

Instructions:

Please submit this problem set before class on Tuesday, October 31. Problem sets may be submitted within a week past the due date at a 20% penalty; each person is allowed to submit one problem late (within a week) without penalty. Please comment your code, because it is part of the requirements of each exercise. Missing comments will not allow the full score.

If you have any questions, please post on the piazza site. This problem set was prepared by Tiziana Sanavia and Giorgio Melloni, so they will be most prepared to answer questions.

1. Random variables and distributions (points: 30)

A. Assume that a die is fair, i.e. if the die is rolled once, the probability of getting each of the six numbers is 1/6. Calculate the probability of the following events.

- Rolling the die once, what is the probability of getting a number less than 4? (points: 5)
- Rolling the die twice, what is the probability that the sum of two rolling numbers is less than 4? (points: 7)

B. Let p be the probability of obtaining a head when flipping a coin. Suppose that Jake flipped the coin n ($n \geq 1$) times. Let X be the total number of head he obtained.

- What distribution does the random variable X follow? Is X a discrete or continuous random variable? (points: 5)
- What is the probability of getting k heads when flipping the coin n times, i.e. what is $Pr(X = k)$ ($0 \leq k \leq n$)? (Write down the mathematical formula for calculating this probability.) (points: 5)
- Suppose p = 0.2 and n = 20. Calculate the probabilities $Pr(X = 4)$ and $Pr(X \geq 4)$. (You may need the functions `dbinom` and `pbinom` in R to calculate these two probabilities. Use `?dbinom` and `?pbinom` to get help information of these two functions). (points: 8)

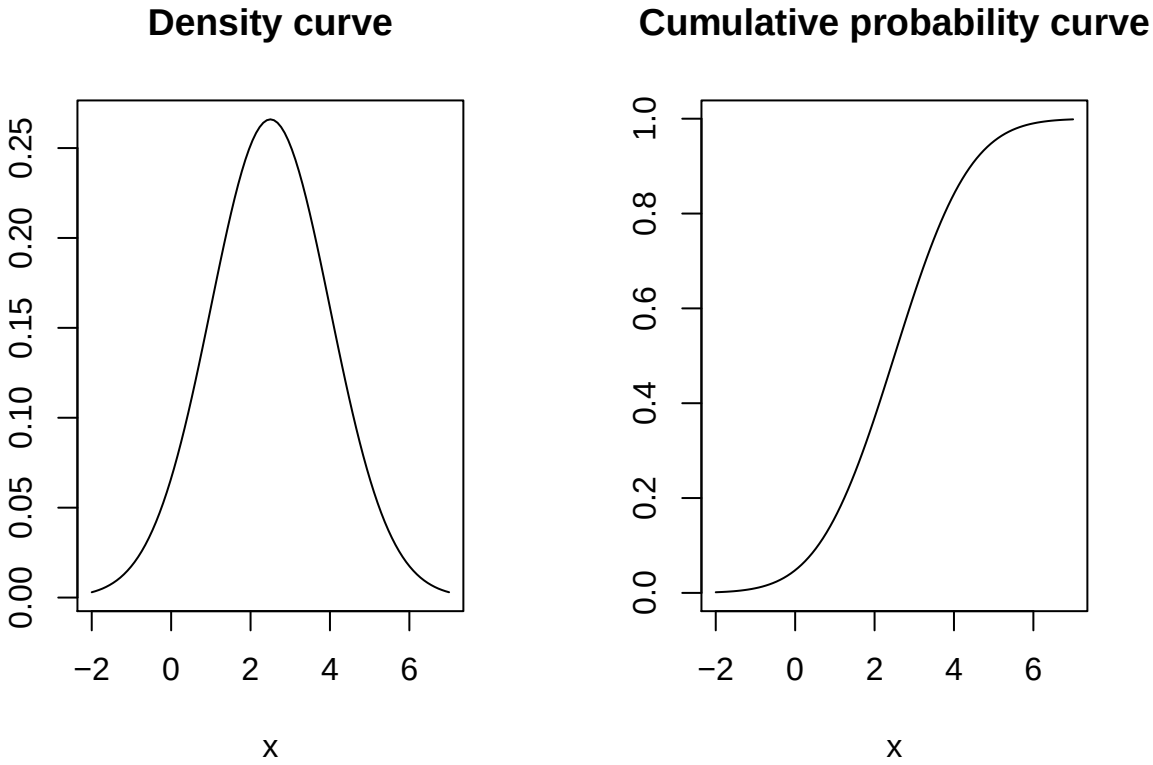
2. Normal Distribution and Z-score (points: 40)

A. The so-called BMI (Body Mass Index) is a measure of the weight-height-relation, and it is defined as the weight (W) in kg divided by the squared height (H) in meters:

$$BMI = \frac{W}{H^2}$$

Assume that the population distribution of BMI is log-normal, therefore log(BMI) is a normal distribution with mean = 2.5 and variance = 2.25.

- Plot in R the density and the cumulative probability curves of log(BMI) as in the picture, using commands `dnorm` and `pnorm`: (points: 4)



- Using the cumulative probability, calculate in R the area under the density curve between x=0.5 and x=4. Use R code for the calculation. (points: 6)
- A definition of "being obese" is a BMI-value of at least 30. How large a proportion of the population would then be obese? (points: 6)
- The 90th percentile of the BMI is the value such that 90% of the population has a BMI lower than this value. Find the 90th percentile for log(BMI) using `qnorm`. (points: 6)

B. Assume that blood-glucose levels in a population of adult women are normally distributed with mean 90 mg/dL and standard deviation 38 mg/dL. Answer the following questions:

- What percentage of women shows levels above or equal to 80.5 mg/dL? (points: 6)
- Suppose that the "abnormal range" is defined to be glucose levels which are 1.5 standard deviations above the mean or 1.5 standard deviations below the mean. What percentage of women would be classified "abnormal"? (points: 6)
- Suppose now that we want to redefine the abnormal range to be more than 'c' standard deviations above the mean or less than 'c' standard deviations with 'c' chosen so that 4 % of the population will be classified as abnormal. What should 'c' be? (points: 6)

3. Simulation of distributions of random variables (points: 30)

Consider X a random variable from any distribution with mean μ and variance σ^2.

If we sample n values from that distribution, we can calculate the mean value $\bar{x_n}$ which is itself the realization of a random variable $\bar{X}_n$.

In this exercise we will evaluate some properties of the:

3.1 Normal Distribution (points: 15)

Using `rnorm` create a vector of 1000 values from a normal distribution with $\mu = 0$ and $\sigma = 1$. We call this vector $m0$. (points: 1)

Using the same command, create a vector of $N = 1000$ mean values from a random sampling of $n = 10, 100$ and 1000 elements. (points: 1) We will call these vectors $m10$, $m100$, $m1000$.

Create a 4 panels plot (You can use an histogram or a density plot or both) showing the distributions of: (points: 2)

1) The 1000 values from the distribution ($m0$)
2) The 1000 means using n = 10 ($m10$)
3) The 1000 means using n = 100 ($m100$)
4) The 1000 means using n = 1000 ($m1000$)

Using the function `qqnorm`, compare theoretical and sample quantiles of a normal distribution. Do the distributions look normal? (points: 3)

Now evaluate the value of the mean and variance of each of the 4 vectors.

- Are the means substantially different from each other? (points: 2)
- Are the variances different from each other? If yes, what is the ratio between $Var(m0)$ and the other variances? (e.g., $Var(m0)/Var(m10)$, $Var(m0)/Var(m100)$) (points: 3)
- If you see any pattern, can you derive a general formula to derive the Variance of any distribution of the means $\bar{X}_n$ for any given n (points: 3)

3.2 Non-normal distribution (points: 15, evaluated as 3.1)

Repeat the exercise 3.1 but using a different random variable following the exponential distribution, $f(x) = \lambda e^{-\lambda x}$. To run this simulation use the function `rexp` with rate (i.e. λ) value **1**.

- Plot the distribution of the 4 vectors
- Using `qqnorm` like above, evaluate normality. Are the exponential values normally distributed? What about the means?
- Evaluate mean and variance as above