Mathematical Statistics - Lab 3

Exercise 1

The distribution of a random variable X is determined by the following probability mass function:

x	**-1**	**0**	**1**	**2**
$P(X = x)$	0.1	0.3	0.1	0.5

a) What is the probability of X being 0.2?
b) Find mean EX and variance VX.
c) $E(X + 6) =$
d) $E(X/2) =$
e) $V(X + 6) =$
f) $V(X/2) =$

Exercise 2

Let X and Y be such independent random variables that $EX = 7$ and $EY = 2$. We also know that $VX = 4$ and $VY = 9$. Find the following:

a) $E(X + Y) =$
b) $E(2X - 5Y) =$
c) $V(X + 3) =$
d) $V(4X - 2Y) =$

Exercise 3

We throw six dice. What is the distribution of X – "number of dice that had a '5' "? How can we write this information down mathematically?

What is the probability that X = 1? What is it for X = 2?

Exercise 4

a) In a town with 10 000 inhabitants, there is an ambulance car that gives first aid and takes people to Tartu's hospital, if necessary (Tartu is quite far away and it takes a day to get there). The probability, that someone needs to be taken to the hospital, is 1/20000 on one day. What is the probability that one ambulance car is not enough (the number of people needing hospitalization exceeds 1)?

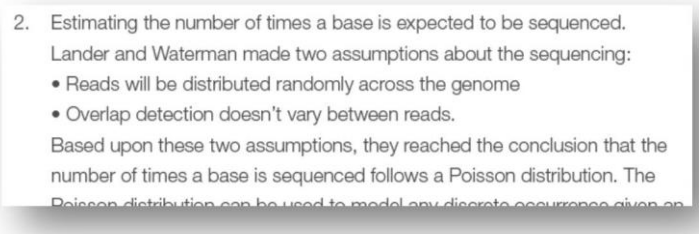
Exercise 5

On average, 7 heart attacks occur in a day in Estonia. The number of heart attacks occurring in a day approximately follows the Poisson distribution. What are the probabilities of 0, 1 and 7 heart attacks happening in one day?

Exercise 6* (for Science and Technology curriculum students (3ECTS))

If we were to find out the genome of a person, then we must count (sequence) some number of letters i.e., base pairs from a starting point that is chosen randomly (for example, we count 250 letters). After that, a random spot is chosen from the genome and another 250 letters is sequenced. These genome sequences are done many times and after that, they are assembled like a puzzle to construct a full genome.

If every spot of a human's genome is covered, on average, with 10 readings, then it is said that the base is read 10 times. The number of times a base is sequenced approximately follows the Poisson distribution.



2. Estimating the number of times a base is expected to be sequenced.
Lander and Waterman made two assumptions about the sequencing:
• Reads will be distributed randomly across the genome
• Overlap detection doesn't vary between reads.
Based upon these two assumptions, they reached the conclusion that the number of times a base is sequenced follows a Poisson distribution. The

(https://www.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf)

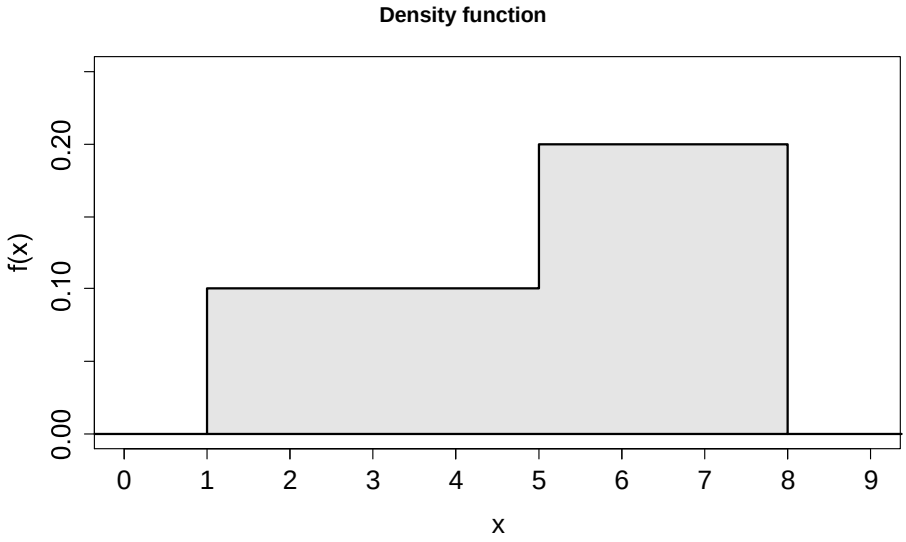
What is the probability that a specific base in the genome is not seen in any readings (read 0 times)?

To think about: the length of human's genome is about 3 000 000 000 letters. How many letters do we expect to not read if we read every letter, on average, 10 times?

Exercise 7

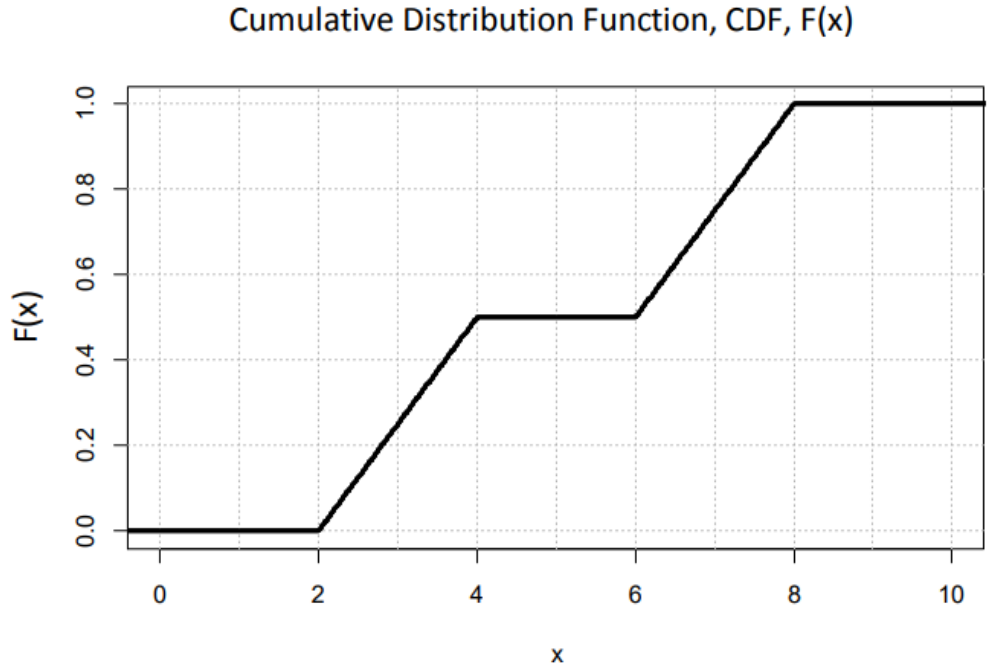
A random variable has a density as described in the plot. Find probability that the random variable's value falls within the interval (4; 6).

$P(4 < X < 6) = \cdots$ 0,1 · 1 + 0,2 · 1



Exercise 8

A random variable has a distribution characterized by the following cumulative distribution function:



Find the following probabilities:

P(X=3) = 0

P(X>6) = $1 - P(X<6)$

$= 1 - 0,5$

P(2<X≤4) = $P(X \leq 4) - P(X \leq 2) = 0,5$

P(4<X≤5) = $P(X \leq 5) - P(X \leq 4) = 0$

Extra exercise*

All the following exercises can be solved using the Binomial distribution – see how you can use this distribution in exercises from real life.

There are dangerous bacteria, and we know a DNA sequence of 60 letters from its genome. It is known that this sequence has not been seen in any other living beings. We read the genome (here we assume that we can read the whole genome all at once) and see, whether we can detect this sequence of 60 letters. If we find it, we have found the dangerous bacteria.

Unfortunately, today's technology makes an error of mistaking one letter while reading the DNA with a probability of 0.01 (these errors are independent events).

a) How many times would we have to read the bacteria's genome, if we want to detect the bacteria with a probability of at least 0.99?

Let's assume that in the sample there is a similar harmless bacterium, with a similar 60-letters-long DNA sequence – with only one letter different.

b) What is the probability of detecting the harmful bacteria in the sample, when, in fact, the sample only contains the harmless bacteria (and the probability of discovering it still 0.99 – we read the genome of the bacteria in the sample as many times we would minimally need to get a probability of discovery of 0.99)?

c) How to achieve a probability of 0.99 for discovering the dangerous bacteria (if the harmful bacteria is, in fact, in the sample) and at the same time make sure that, if only the harmless bacteria is in the sample, the probability of a false positive test result would be less than 0.01?

The answers to the extra exercise (try to solve it yourself first before peeking!):

a)6 or more times
b)0,019834…
c)If we were to read an unknown bacteria's genome 9 times and insist that we see the seeked DNA sequence of 60 letters two or more times.
If the bacteria in the sample were the harmful one, we would detect it with a probability of ~99.05%, but if the sample contains the harmless bacteria, then we would say it's the harmful one with a probability of ~0.0004.
There are other possible combinations for solving this exercise (e.g. reading a sample 12 times and insisting that we see the dangerous bacteria's signature 3 or more times).

4 $X \simeq$ number of people sick in 1 day

1) $X \sim Bi\left(10000, \frac{1}{20000}\right)$

2) $X \sim Po\left(\frac{10000}{20000}\right)$

$$P(X>1) = 1 - P(0) - P(1) =$$

$$= 1 - 1 \cdot \left(\frac{1}{20000}\right)^0 \left(\frac{19999}{20000}\right)^{10000} -$$

$$- 10000 \left(\frac{1}{20000}\right)^1 \left(\frac{19999}{20000}\right)^{9999} =$$

$$= 1 - 0{,}607 - 0{,}5 \cdot 0{,}6075 \simeq 0{,}0895$$