**Statistics and data analysis 2018**

**Final Exam (Alef)**

Guidelines

* There are **4** (**FOUR**) questions in the exam. You need to answer **all** of them (no choice).
* You can respond in English and/or Hebrew.
* Write the answers to the questions in the exam notebook.
* Justify all your answers. Even though many of the questions are not purely mathematical, you should mathematically explain your answers. You may assume results proven (or stated as a fact) in class or in the homework (unless the question instructs otherwise).
* Make sure you write in a clear and legible way. Grading will also depend on the clarity and not only on correctness.
* You can use the reference and formulae sheet as provided, including the standard normal table. No other auxiliary material can be used during the exam.
* The total time of the exam is 3 (three) hours.
* Good luck!

Question 1 (25 pts)

1. (10 pts)  
   Consider the pairs of observed measurements depicted in the next page. There are five of them.   
   Determine a matching between Pearson and Spearman correlation values in the rows of Table 1 below and the letter enumeration (A to E in Fig 1) of the depicted cases.  
   Indicate the matching clearly in your notebook.

Table 1:

|  |  |  |
| --- | --- | --- |
| Number (to be matched to the figures) | Pearson correlation | Spearman correlation |
| 1 | 0.87 | 0.63 |
| 2 | -0.2 | 0.5 |
| 3 | 0.96 | 0.63 |
| 4 | 0.89 | 0.89 |
| 5 | 0.2 | -0.5 |

Fig 1 (A-E):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| A |  | | B |  | |
| C |  | | D |  | |
| E | |  | | |

1. (15 pts)
   1. (5 pts) Given the following 2 datasets:
      1. 
      2. 

You need to report the statistical significance of the difference between circles and crosses.   
What will be the difference between your report for i and your report for ii if you use a one sided Student t-test?  
What will be the difference between your report for i and your report for ii if you use WRS?   
Justify your answers.

* 1. (10 pts) Given the following 2 datasets:
     1. 
     2. 

You need to report the statistical significance of the difference between circles and crosses.   
What will be the difference between your report for i and your report for ii if you use a one sided Student t-test?  
What will be the difference between your report for i and your report for ii if you use WRS?   
Justify your answers.

Calculate the WRS in each of the datasets (show the calculation only – no need to provide a final numerical result).    
Can you calculate the WRS p-value in any of these cases?

Question 2 (25 pts)

1. (12 pts) Define two random variables X and Y that assume values on the non-negative integers so that:
   * Both X and Y assume at least two values with non-zero probability (they are not constant)
   * The random variable Z = X+Y is uniformly distributed over the numbers {10, 12, …, 134} (all even numbers between 10 and 134, inclusive)
2. (3 pts) Do section A again, providing random variables X and Y that satisfy the same conditions, but are different from the ones you defined above.
3. (10 pts)  
   Let X be a random variable. (X has Poisson distribution with ).
4. (5 pts) What is ? (no need to calculate exact numbers, just provide a clear mathematical expression)
5. (5 pts) True or False?

Explain your answer.

\* is the entropy of X.

Question 3 (25 pts)

This question has 5 parts numbered A-E.  
  
A scientist is generating nanoparticles for an experiment. She observes the following distribution of particle radii, in nms (nano-meters):  
  


This histogram representation of the distribution is calculated from 100K particles. The x-axis units are nms. The histogram is truncated at 20 nm. 30687 particles of the 100K measured had radius ≥ 20 nm.

1. (5 pts)   
   For the above data representing 100K particles, the scientist calculated empirical statistics.   
   The empirical mean of the data is nm   
   The empirical standard deviation is nm.   
   The empirical median of the distribution is at nm.   
   Let R denote the random variable that represents radii of the particles generated by the scientist.   
   What do you think the distribution of R is? Explain your answer.
2. (5 pts)   
   The scientist produced QQ plots for her data against the standard normal distribution. Amongst the 4 plots (marked A-D) in the next two pages, indicate which one (if any) corresponds to a QQ-plot of the quantiles of log(R) and which one (if any) corresponds to a QQ-plot of the quantiles of R. Explain your answer.

  
  
  


B

A

C

D

1. (5 pts)   
   According to the model you have developed what is the radius r so that   
   # of particles with radius < r = 20000? (leave answer in exp notation if necessary)
2. (5 pts)   
   The experiment requires at most 10% of particles to have a radius larger than nm. Show, based on your model, that the population generated here is therefore not adequate for the experiment.
3. (5 pts)   
   The scientist can treat the particles and decrease all particle radii.   
   A reasonably priced process will lead to all radii decreasing exactly fold (a particle with radius will have radius after the treatment).   
   A more expensive process will lead to all radii decreasing exactly fold (a particle with radius will have radius after the treatment).   
   She consulted with her statistician colleague as to whether either of the treatments will solve the problem and specifically as to whether the less expensive one will do it.   
   What advice would you give in this case? Show all your calculations.

Question 4 (25 pts)

In this question stands for a normal distribution with mean and standard deviation .

Fred, Mel and Sid are repair technicians who work for Randobezeq – a phone company.

Fast Fred takes time which is N(30,25) to repair a telephone line failure at a customer’s home.

Medium Mel takes time which is N(35,49) for the same task.

Slow Sid takes time which is N(40,100) for the same task.

1. (5 pts) Fred is due to arrive to repair your phone at 10AM tomorrow. How confident can you be that you will be done by 10:38?
2. When a customer in North Randomistan orders a repair, there is a 40% chance Fred will do the work and 30% each that Mel or Sid will do the work.
   1. (2 pts) What is the distribution of the duration of repair in North Randomistan?
   2. (8 pts) Let Φ denote the CDF of a standard normal random variable. Use Φ to express the CDF of the duration of a repair in North Randomistan. Explain your answer.
   3. (10 pts) If the repair starts at 10AM, which of the following times is the earliest time by which the customer can assume, with a 95% certainty, that the repair will already be done?  
      State only one of the following options in your notebook and then justify and explain your answer (you may use the formula you developed above).   
      Options:  
        
      10:23  
      10:36  
      10:42

10:51  
11:04