**Statistics and data analysis 2018**

**Final Exam (Alef) - Solution**

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

**Solution**

Question 1 (25 pts)

|  |  |  |  |
| --- | --- | --- | --- |
| Number (to be matched to the figures) | Pearson correlation | Spearman correlation | Correct Plot |
| 1 | 0.87 | 0.63 | B |
| 2 | -0.2 | 0.5 | D |
| 3 | 0.96 | 0.63 | A |
| 4 | 0.89 | 0.89 | E |
| 5 | 0.2 | -0.5 | C |

* 1. t-test: ii will have a more significant p-value. The reason is that the means of the two class samples (circles vs crosses) in ii are far from each other comparing to i.

WRS: there will be no difference in the WRS, since WRS is a rank test, and the ranks (of circles, say) in both cases are the same.

* 1. t-test: ii will have a more significant p-value. The reason for that is, although the means in both cases are the same, the empirical variances are different and smaller in ii.   
     WRS: there will be no difference in the WRS, as above.

WRS calculation:

To = 1+2+3+4+5 = 15, Tx = 6+7+8+9+10 = 40

We can calculate the p-value using combinatorial approach: .

The normal approximation is not good enough in this small dataset (the difference in the p-value is about 50%). It is a reasonable approach, though, and we did give credit for solutions based on normal approximation.

Question 2 (25 pts)

1. X = [0, 42, 84], p=1/3

Y = [10, 12, …, 50], p=1/21

1. X = [0, 18, 36, 54, 72, 90, 108], p=1/7

Y = [10, 12, …, 26], p=1/9

\* There are many more options



1. True

Since   
we have

Question 3 (25 pts)

1. The distribution of R is log-normal. We first can see it from the histogram. We can take the log-normal median , where is the mean of the underlying normal distribution, and deduce that the parameters of the distribution are , by solving .  
   The parameters should also be tested against the given observed empirical parameters.
2. The quantiles of R correspond to plot B – this is how log-normal distribution looks like in QQ plot (and also this is the only plot with only positive values in the y-axis).

The quantiles of log(R) correspond to plot D – if R has log-normal distribution it means that , where X has normal distribution, –> –> log(R) has normal distribution, and D plot is the QQ of normal distribution.

1. We want to find r, such that, 20% of the distribution have radius less than r. In the standard normal this corresponds to Φ-1(0.2) = − 0.84 stds (using the table).   
   Therefore:
2. in the log-normal distribution corresponds to 4 in the underlying normal distribution. We now need to find the z-score for the point 4:

This means that about 15.9% of the particles have radius larger than .

1. Expensive procedure:

Note that if Y is LogNormal then cY is also LogNormal, for any positive scalar c.

In our case, the median before the procedure was , so now it will be the mean of the underlying normal distribution is 1.

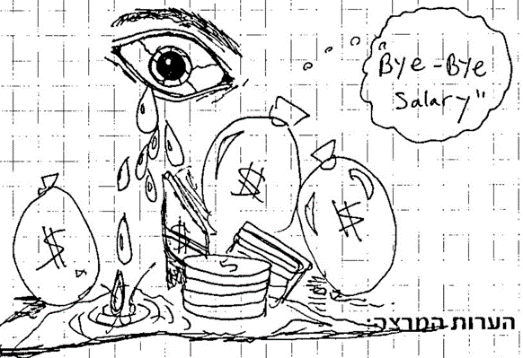
The mean before the procedure was , so now it will get

Now we can do the same calculation as in section d:

The expensive procedure achieves the goal.

Cheaper procedure:

The median before the procedure was , so now it will be the mean of the underlying normal distribution is 1.5.

The mean before the procedure was , so now it will get

Now can do the same calculation as in section d:

The cheaper procedure doesn’t achieve the goal.

NIce ... Bye bye Research Budget, actually ... :)

Question 4 (25 pts)

1. 1. The distribution of the duration of repair in North Randomistan is Gaussian Mixture with the following parameters:
   2. In order to solve this question, we can try all the time possibilities. If we want to solve it more efficiently, we can conclude from section 1 that the first 2 times will give us less than 95% certainty. In addition, we can start from the middle and go to the correct direction according to the calculated result (Binary Search).

10:42

In 10:42 we have less than 95% certainty.

10:51

In 10:51 we have more than 95% certainty.