**Statistics and data analysis 2019**

**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 exam notebooks. Don’t use the exam printout.
* 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.
* Use normal approximation when appropriate and needed.
* You can use hand held calculators.
* 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. (6 pts)  
   Consider the pairs of observed measurements below. There are three of them.   
   Determine a matching between Pearson and Spearman correlation values in the rows of Table 1 below and the letter enumeration (A to C 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 | 0.93 | 0.87 |
| 2 | 0.98 | 1 |
| 3 | 0.94 | 0.9 |

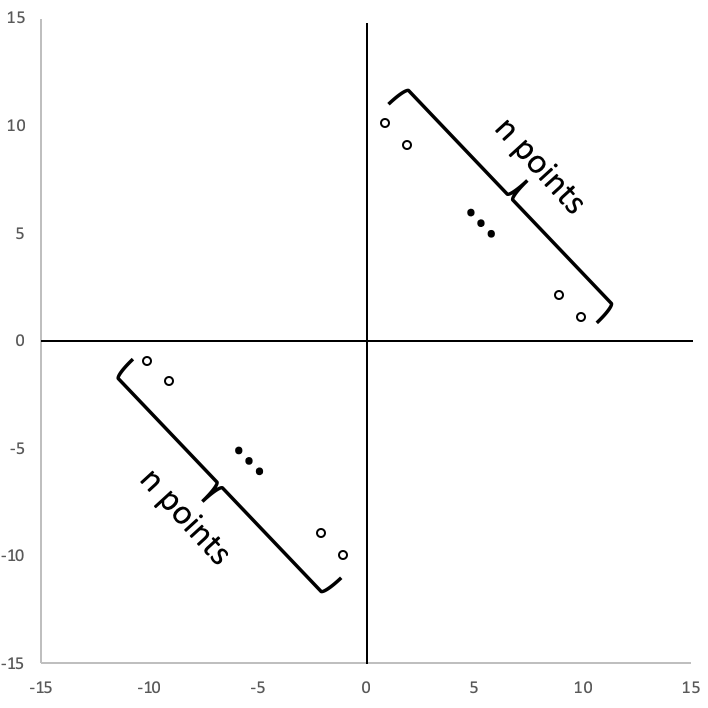
Fig 1 (A-C):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| A |  | | B |  | |
| C | |  | | |

1. (4 pts) Can you add one data point to the following dataset so that:
   1. Pearson correlation will be positive?
   2. Spearman correlation will be positive?

Justify you answer.

1. (5 pts) Consider the following dataset D(n), defined by the following picture:



Let = Kendall correlation of the dataset D(n).

Find:

Prove your answer.

1. (10 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, in both scenari.   
For this purpose, as a first approach, you are computing the WRS statistics and the associated p-values.

Write an expression for the p-value you will get for D1 and for D2.  
Justify your answers.  
As a second approach you are using a t-test. Your null hypothesis is that the crosses and the circles come from two distributions with the same mean. For which of the two datasets will you reject the null with more confidence?  
Explain your answer.

* 1. (5 pts) Same as above for the following 2 datasets:
     1. 
     2. 

Question 2 (25 pts)

1. (5 pts) Recall the coupon collector scenario described in class, where we have n countries, and each has equal probability for the next visit in the website and where every visit is independent of all previous visits.

Let Xi = the number of visits, after the first i-1 countries are in, until the i-th country is also in.

What is the distribution of Xi? Explain.

1. (10 pts) Let n=3 (3 countries) and T=X1+X2+X3 (Xi the same as in section a), represent the time it takes to have seen all countries.

Calculate the distribution of T in the range . That is – compute P(T=j) for all js in the indicated range.

1. (5 pts)  
   Consider the following confidence intervals generated from a Bernoulli distribution with the same n and with different values of alpha.

In your notebook state the order of the outcomes according to the size of alpha (low alpha – high confidence, to high alpha – low confidence). Justify your answer.

|  |  |  |
| --- | --- | --- |
| A | B | C |
|  |  |  |

1. (5 pts)  
   Consider the following confidence intervals generated from a Bernoulli distribution with different ns and with the same value of alpha.

In your notebook state the order of the outcomes according to the size of n (low to high). Justify your answer.

|  |  |  |
| --- | --- | --- |
| A | B | C |
|  |  |  |

Question 3 (25 pts)

In this question stands for a Poisson distribution with mean .

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

Fast Fred takes time which is to repair a telephone line failure at a customer’s home.

Slow Sid takes time which is 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 he will be done by 10:05?
2. (5 pts) Given 2 Poisson distributions with the parameters and 2 mixture coefficients , we define a Poisson mixture distribution, M, by writing its PDF:

Prove that:

1. When a customer in North Randomistan orders a repair, then the distribution of the repair time is a Poisson mixture with mean = 16.
   1. (5 pts) What is the probability that following a call in North Randomistan Fred is providing the service?
   2. (10 pts) Under the condition of this question we can calculate that Fred completes 99% of the cases in 18 minutes or less and Sid completes 56% of the cases in 18 minutes or less.

If a repair in North Randomistan starts at 10AM, which of the following times is the earliest time by which the customer can assume, with a 50% 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.   
Options:  
  
10:08  
10:10  
10:13

10:16

10:18

10:21

Question 4 (25 pts)

A survey was conducted in two Randomistan Farms, farm A and Farm B as to parameters that may affect the susceptibility of apple trees to fungus.

In Farm A the survey covered k(A) = 40 trees, Aff(A) = 15 of them were affected and   
N(A) = 25 of them were unaffected.

In Farm B the survey covered k(B) = 90 trees, Aff(B) = 15 of them were affected and N(B) = 75 of them were unaffected.

The survey measured 100 different features of the trees and sought to determine features that are associated with higher susceptibility.

1. (8pts) The survey found that the height of the tree is associated to susceptibility. Ranking trees from tallest to shortest, denote the sum of ranks of affected trees in Farm A and in Farm B by RS(A) and RS(B) respectively.   
   The survey found that RS(A) = 160 and RS(B) = 200. For which of the farms do we have a more significant p-value to support the stated statistical association? Justify your answer and show calculations that support it.
2. (7 pts) In Farm A the survey yielded 20 features at an FDR of 0.05 and 40 features at an FDR of 0.1.  
   In Farm B the survey yielded 10 features at an FDR of 0.05 and 40 features at an FDR of 0.1.  
   In your notebook draw possible graphs that describe p-values of features (x-axis) with the number of observed features Obs(x) (y axis), one graph for each one of the two farms.
3. (5 pts) Let X~Binom(0.5, 4). Compute the entropy H(X).
4. (5 pts) Let Y~Binom(0.5, 15). True or false: H(Y) < 4?