Statistics and data analysis 2019

Final Exam (Gimel)

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!

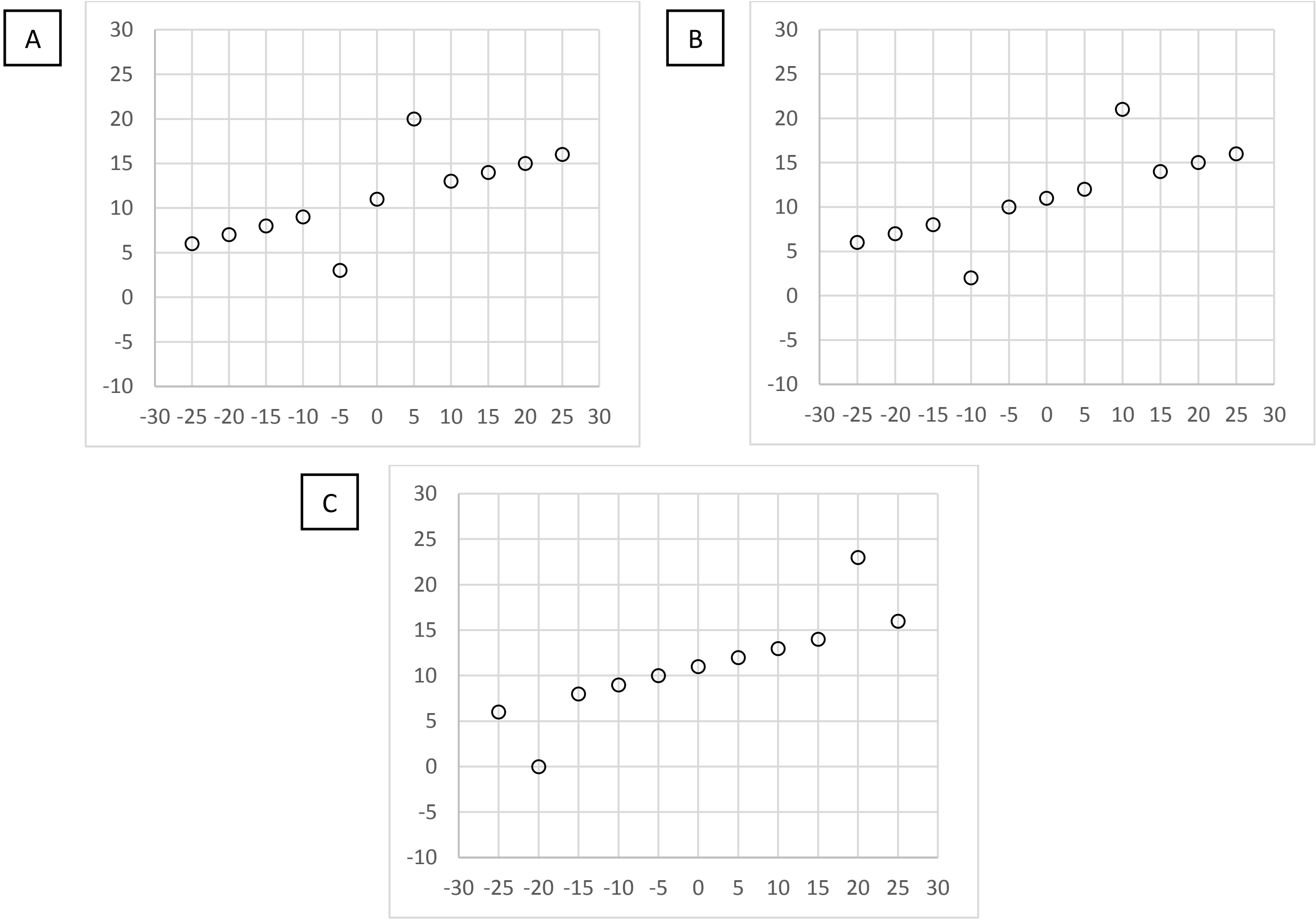
Question 1 (25 pts)

A. (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.

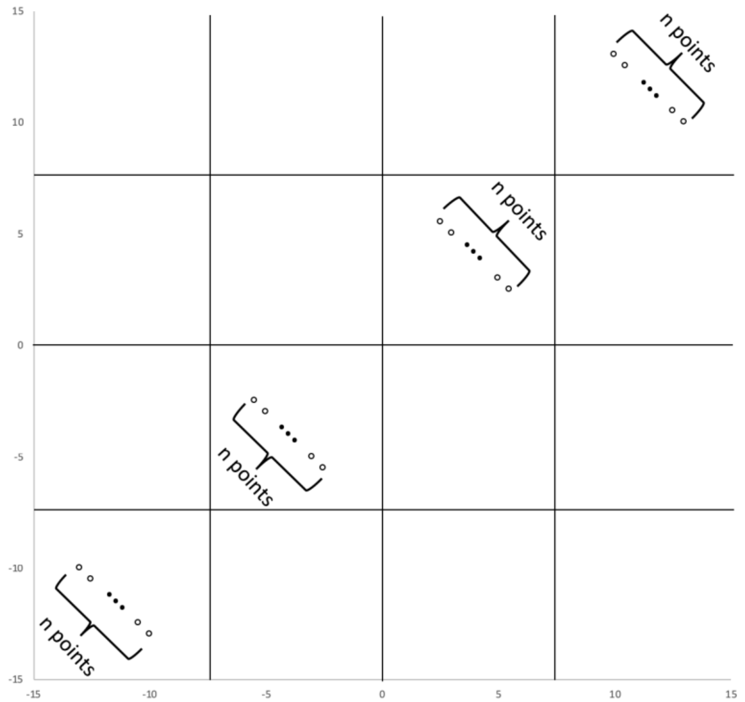
Table 1:

Number (to be matched to the figures)	Pearson correlation	Spearman correlation
1	0.87	0.98
2	0.75	0.82
3	0.79	0.89

Fig 1 (A-C):



B. (7 pts) Consider the following dataset D(n), defined by the following picture:



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

Find:

$$\lim_{n\to\infty} \tau(n)$$

Prove your answer.

C. (12 pts)

Safety tests were conducted for cars made in the Randomistan Opel factory and in the Germany Opel factory. Higher safety scores are better. The house statistician in the company had decided to declare the Randomistan cars safer if the WRS p-value of the observed data is better than 0.15.

For each one of the situations described below state whether Randomistan cars are declared safer.

In #3 also state what would happen if one were to use Student t-test rather than a WRS test.

Explain your calculations and answers.

1. (3 pts)

Randomistan	9.3	8.8	8.5		
Germany	9.1	8.2	8.1	8	7.9

2. (3 pts)

Randomistan	8.8	8.6	8.1		
Germany	9.2	9.1	9	8.9	5

3. (6 pts)

Randomistan	10	0.06	0.05					
Germany	9.95	9.9	9.85	0.04	0.03	0.02	0.01	0

Question 2 (25 pts)

A.

1. (5 pts) Define two random variables X and Y that assume values on the non-negative integers so that:
    a. Both X and Y assume at least two values with non-zero probability (they are not constant)
    b. Let Z = X+Y. Then Z is uniformly distributed over the numbers {5, 6, 7, … , 114, 115}.
2. (5 pts) H(X) < 5? True or False? Explain.
3. (5 pts) H(Y) < 5? True or False? Explain.

B. Recall that the negative binomial distribution NegBinom(p,r) represents the number of times a coin with P(H) = p is tossed until the first time we see exactly r Hs.

1. (4 pts) Explain why if T ~ NegBinom(p,r) then

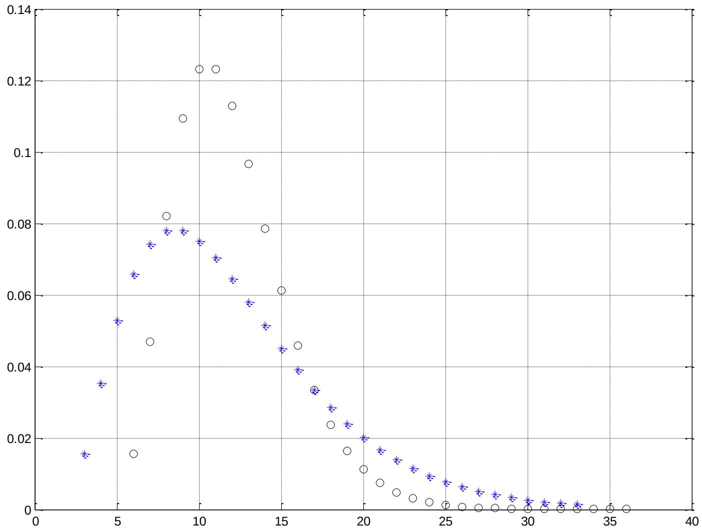
    $$P(T = t) = \binom{t-1}{r-1} p^r (1-p)^{t-r}$$

2. (3 pts) In the following figure we present the pdfs of two negative binomial distributions.

    One with r1 = 3 and p1 = ??

    The other with r2 = 6 and p2 = 0.5.

    Determine which is which (and explain your answer)



3. (3 pts) Now assume that r1=6, r2=12 and p2=0.5. Further assume that the two distributions have the same mean what is p1?

Question 3 (25 pts)

In this question $Poi(\lambda)$ stands for a Poisson distribution with mean λ.

Fred and Sid are repair technicians who work for Randobezeq – a phone company.
Fast Fred takes time which is $Poi(10)$ to repair a telephone line failure at a customer's home.
Slow Sid takes time which is $Poi(18)$ for the same task.

A. (5 pts) Fred is due to arrive to repair your phone at 10:05AM tomorrow. How confident can you be that he will be done by 10:10?

B. (5 pts) Given 2 Poisson distributions with the parameters λ_1, λ_2 and 2 mixture coefficients w_1, w_2, we define a Poisson mixture distribution, M, by writing its PDF:

$$P(i) = w_1 P_1(i) + w_2 P_2(i)$$

Prove that:

$$E(M) = w_1 \lambda_1 + w_2 \lambda_2$$

C. 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 the repair starts at 10:05AM, 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:

   10:13
   10:15
   10:18
   10:21
   10:23
   10:26

Question 4 (25 pts)

A.

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.

B.

A bike store in Randomistan sells bike parts and supplies including a popular chain oil called Mighty Oil. The store starts every morning with a stock of 40 bottle.
The store collected data and determined that the daily demand is normally distributed with a mean of μ=20 bottles and a standard deviation of σ = 5 bottles.
The store works 300 days annually.

1. (5 pts) In how many days, annually, should the store manager expect to get, during the day, to a stock of 0?
   Such a day is called a stock-out day.
2. (5 pts) Two years later the sales increased to be normally distributed with a mean of μ=30 bottles and a standard deviation of σ = 6 bottles. The store manager asked for your statistical advice regarding the changing in the stock in the beginning of the day.
   What should that number be to results in 3 stock-out days per year? 6 stock-out days per year?