Statistics and data analysis 2019

Final Exam (Alef) – with 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.
- 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)

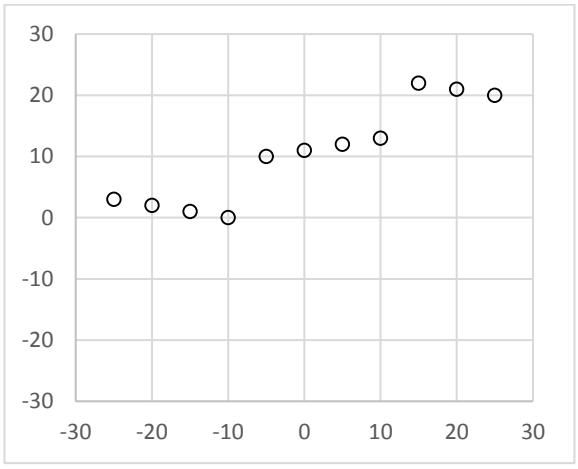
Consider the pairs of observed measurements below. There are three of them. Determine a matching between Pearson and Spearman correlation values in the rows of Table 1 below and the letter enumeration (A to C in Fig 1) of the depicted cases. Indicate the matching clearly in your notebook.

Table 1:

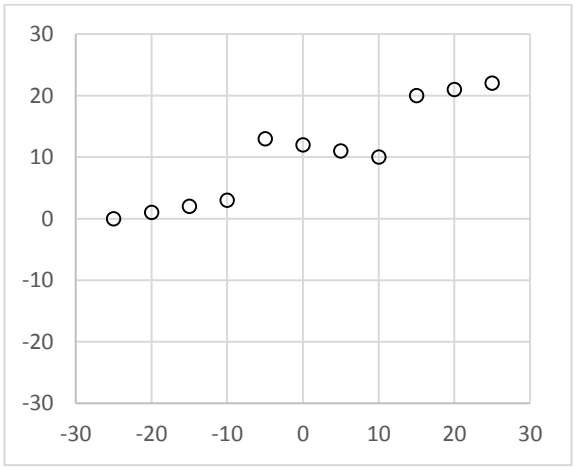
Number (to be matched to the figures)	Pearson correlation	Spearman correlation
1	0.93	0.87
2	0.98	1
3	0.94	0.9

Fig 1 (A-C):

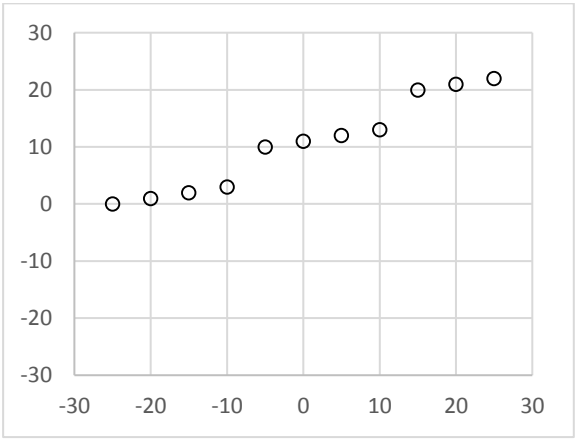
A



B



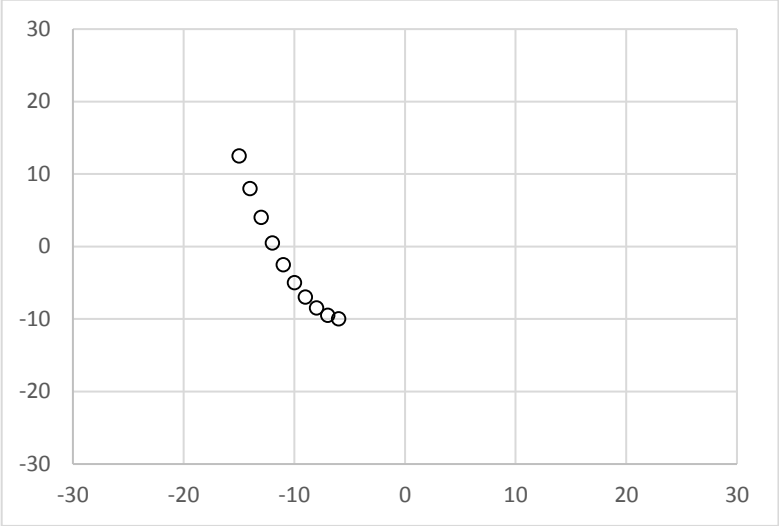
C



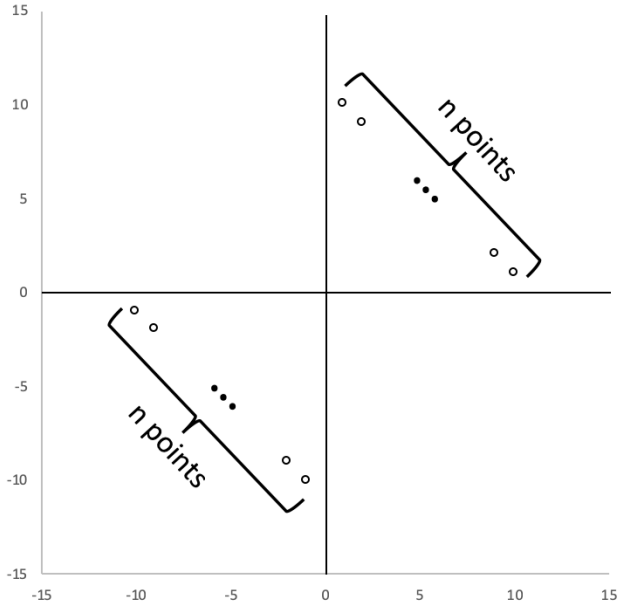
B. (4 pts) Can you add one data point to the following dataset so that:

1. Pearson correlation will be positive?
2. Spearman correlation will be positive?

Justify you answer.



C. (5 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.

D. (10 pts)

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

D1.

D2.

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 two-sided p-values.

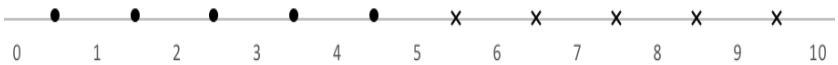
Write an expression for the p-value you will get for D1 and for D2.

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.

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

D1.



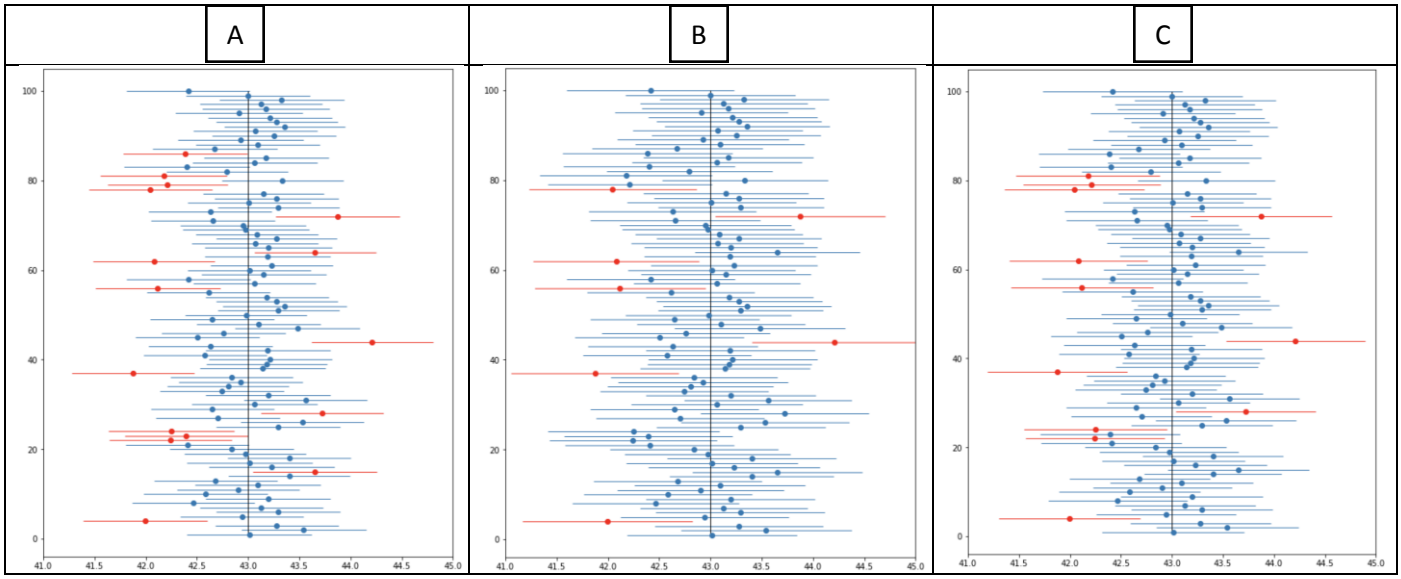
D2.

Question 2 (25 pts)

A.

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.
2. (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 $1 \leq j \leq 6$. That is – compute P(T=j) for all js in the indicated range.

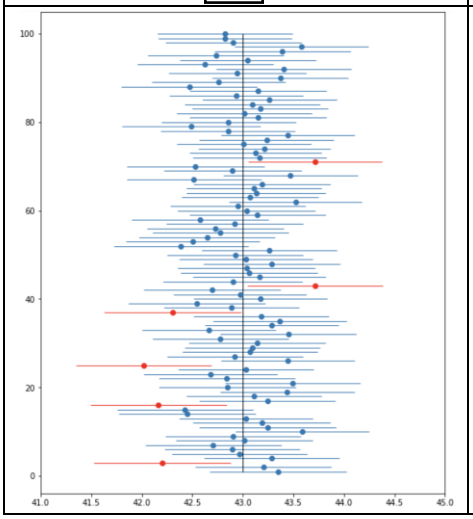
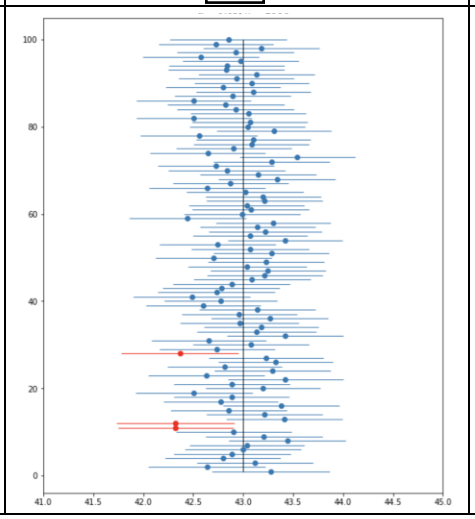
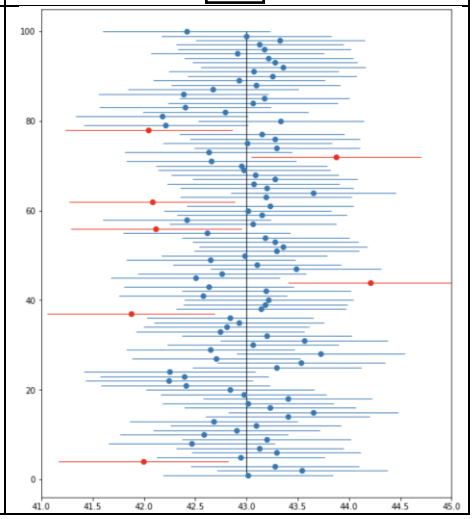
B. (5 pts)
Consider the following confidence intervals generated from a Bernoulli distribution with the same n and with different values of alpha.
In your notebook state the order of the outcomes according to the size of alpha (low alpha – high confidence, to high alpha – low confidence). Justify your answer.



C. (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 $Poi(\lambda)$ 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 10AM tomorrow. How confident can you be that he will be done by 10:05?
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 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.

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 <u>affected trees</u> 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.

1. (5 pts) Let X~Binom(0.5, 4). Compute the entropy H(X).
2. (5 pts) Let Y~Binom(0.5, 15). True or false: H(Y) < 4?

Solution

Question 1 (25 pts)

A.

Number (to be matched to the figures)	Pearson correlation	Spearman correlation	Correct Plot
1	0.93	0.87	A
2	0.98	1	C
3	0.94	0.9	B

B.

1. Yes. Because Pearson is sensitive to outlier we can add a point (x, y) at the upper right corner. If the (x, y) values are large enough then the Pearson correlation will be positive.
2. No. Spearman is looking at the ranks of the points and not at the values. Therefore, one point won't change 10 monotonically decreasing ranks.

C. $\tau = \frac{C-D}{\binom{n}{2}}$.

C is the number of concordant pairs – in this case n^2.

D is the number of discordant pairs – in this case $\binom{2n}{2} - n^2 = n^2 - n$

$$\tau(n) = \frac{n^2 - n^2 + n}{\binom{2n}{2}} = \frac{n}{n(2n-1)} = \frac{1}{2n-1} \xrightarrow[n\to\infty]{} 0$$

D.

1. For the dataset D1 let T=WRS(circles)
   We first calculate T=21.
   We then note that E(T)=5*6=30. Therefore, we observed a smaller WRS than expected at random.
   We now use normal approximation to compute the p-value for this observation:

$$\sigma_T = \sqrt{\frac{5 * 6 * 12}{12}} = \sqrt{30}$$

   For the dataset D1 we therefore get:

$$Z(T) = \frac{T - E(T)}{\sigma_T} = \frac{21 - 30}{\sqrt{30}} = -1.643$$

   We know that Z(T) is approximately standard normal. Using the z-score table we get:

$$P(T \leq 21) = P(Z(T) \leq -1.643) \approx 1 - 0.95 = 0.05$$

   We now apply the same process to the dataset D2.

T=20.
E(T)=5*5.5=27.5.

$$\sigma_T = \sqrt{\frac{5*5*11}{12}} = \sqrt{22.92} = 4.787$$

$$Z(T) = \frac{T - E(T)}{\sigma_T} = \frac{20 - 27.5}{4.787} = -1.567$$

$$P(T \leq 20) = P(Z(T) \leq -1.567) \approx 1 - 0.941 = 0.059$$

For the t-test we will get a smaller p-value in D2 because the empirical averages are farther from each other. Therefore, in D2 we will reject the null hypothesis with higher confidence.
For D2 we have 10 points vs 11 points for D1, but this difference is not sufficient to offset the difference in $\mu(circles) - \mu(crosses)$.

2. Again let T=WRS(circles).
   For D1 note that T=15, which is the smallest possible value for this configuration. The p-value is:

$$P(T \leq 15) = \frac{1}{\binom{10}{5}}$$

   For D2 note that T=36 and:

$$P(T \leq 36) = \frac{1}{\binom{16}{8}}$$

   Working with t-test we note that the distance between the averages is slightly larger in D2. Moreover, the empirical std is smaller in D2 and the N is larger. All of this implies that we reject the null hypothesis in higher confidence in D2.

Question 2 (25 pts)

A.

1. The distribution of Xi is Geometric with $P = \frac{n-i+1}{n}$.
   This is because after i-1 countries we are waiting for one of the other n-i+1 to visit the site. Success will be a visit of one of the n-i+1 countries and failure will be a visit of one of the i-1 countries that are already in. This is exactly the Geometric distribution when we are waiting for the first success. The probability for success will be the probability for new country to visit the site which is.
   $P = \frac{n-i+1}{n}$ since all countries have the same probability 1/n.
2. First, we note that:
   $X_1 \equiv 1$
   $X_2 \sim Geo\left(\frac{2}{3}\right)$
   $X_3 \sim Geo\left(\frac{1}{3}\right)$
   We explicitly write the distributions for i=1-6:

	1	2	3	4	5	6
X_1	1	0	0	0	0	0
X_2	$\frac{2}{3}$	$\frac{1}{3}\cdot\frac{2}{3}$	$\left(\frac{1}{3}\right)^2\cdot\frac{2}{3}$	$\left(\frac{1}{3}\right)^3\cdot\frac{2}{3}$	$\left(\frac{1}{3}\right)^4\cdot\frac{2}{3}$	$\left(\frac{1}{3}\right)^5\cdot\frac{2}{3}$
X_3	$\frac{1}{3}$	$\frac{2}{3}\cdot\frac{1}{3}$	$\left(\frac{2}{3}\right)^2\cdot\frac{1}{3}$	$\left(\frac{2}{3}\right)^3\cdot\frac{1}{3}$	$\left(\frac{2}{3}\right)^4\cdot\frac{1}{3}$	$\left(\frac{2}{3}\right)^5\cdot\frac{1}{3}$

   Let $S = X_1 + X_2$
   The distribution of S is clearly a shift by 1 of the distribution of X_2:

	1	2	3	4	5	6
S	0	$\frac{2}{3}$	$\frac{1}{3}\cdot\frac{2}{3}$	$\left(\frac{1}{3}\right)^2\cdot\frac{2}{3}$	$\left(\frac{1}{3}\right)^3\cdot\frac{2}{3}$	$\left(\frac{1}{3}\right)^4\cdot\frac{2}{3}$

   To get the distribution of T we use convolution of S and X_3, which are independent under our conditions.
   $P(T=1) = 0$
   $P(T=2) = 0$
   $P(T=3) = P(S=2) * P(X_3=1) = \frac{2}{3} * \frac{1}{3} = \frac{2}{9}$
   $P(T=4) = P(S=2) * P(X_3=2) + P(S=3) * P(X_3=1)$
   $= \left(\frac{2}{3}\right)^2 \cdot \frac{1}{3} + \left(\frac{1}{3}\right)^2 \cdot \frac{2}{3} = \frac{6}{27} = \frac{2}{9}$
   $P(T=5) = P(S=2) * P(X_3=3) + P(S=3) * P(X_3=2)$
   $+ P(S=4) * P(X_3=1) = \left(\frac{2}{3}\right)^3 \cdot \frac{1}{3} + \left(\frac{2}{3}\right)^2 \cdot \left(\frac{1}{3}\right)^2 + \frac{2}{3} \cdot \left(\frac{1}{3}\right)^3 = \frac{14}{81}$
   $P(T=6) = P(S=2) * P(X_3=4) + P(S=3) * P(X_3=3)$

$$+ P(S = 4) * P(X_3 = 2) + P(S = 5) * P(X_3 = 1)$$

$$= \left(\frac{2}{3}\right)^4 \cdot \frac{1}{3} + \left(\frac{2}{3}\right)^3 \cdot \left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 \cdot \left(\frac{1}{3}\right)^3 + \frac{2}{3} \cdot \left(\frac{1}{3}\right)^4 = \frac{30}{243} = \frac{10}{81}$$

B. The order is $\alpha(B) < \alpha(C) < \alpha(A)$.
We can see that in B we have the largest intervals and therefore the smallest number of intervals that don't cover the mean.
In A we can see exactly the opposite.
Note that the centers of the intervals are the same in all three plots and the only difference is the size of the intervals. Which for fixed n is inversely monotonic in alpha.

C. The order is $n(C) < n(A) < n(B)$.
We can see that in B we have the smallest intervals which means the largest n.
In C we can see exactly the opposite.
Note that the number of intervals that don't cover the mean is not exactly the same as you would expect with a fix alpha. This is due to a random variation.

Question 3 (25 pts)

A. $P(X \le 5) = P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5) =$

$$= \sum_{k=1}^{5} \frac{e^{-10}10^k}{k!} = 0.067 \to 6.7\%$$

B.

$$E(M) = \int_{-\infty}^{\infty} xP(x) \overset{\text{In Poisson}}{=} \sum_{0}^{\infty} xP(x) =$$

$$= \sum_{0}^{\infty} x\big(w_1P_1(i) + w_2P_2(i)\big) = \sum_{0}^{\infty} xw_1P_1(i) + xw_2P_2(i) =$$

$$= \sum_{0}^{\infty} xw_1P_1(i) + \sum_{0}^{\infty} xw_2P_2(i) = w_1\sum_{0}^{\infty} xP_1(i) + w_2\sum_{0}^{\infty} xP_2(i) =$$

$$= w_1\lambda_1 + w_2\lambda_2$$

The last step follows from the fact that when $X \sim Poi(\lambda)$ then $E(X) = \lambda$

Q.E.D

C.

1. From B we know that:

$$E(M) = w_1\lambda_1 + w_2\lambda_2 \to 16 = w_1 10 + w_2 18$$

We also know that:

$$w_1 + w_2 = 1 \to w_2 = 1 - w_1$$

And we get:

$$16 = w_1 10 + (1 - w_1)18 = 18 - 8w_1$$

$$w_1 = 0.25 \to 25\% \quad (w_2 = 0.75)$$

2. From CDF definition we have:

$$CDF(18) = w_1CDF_1(18) + w_2CDF_2(18)$$

From the question details:

$$CDF(18) = 0.25 * 0.99 + 0.75 * 0.56 = 0.6675$$

Now, In order to calculate CDF(17), we can use the result of CDF(18):

$$CDF(17) = CDF(18) - w_1PDF_1(18) + w_2PDF_2(18) =$$

$$= 0.6675 - 0.25\frac{e^{-10}10^{18}}{18!} - 0.75\frac{e^{-18}18^{18}}{18!} = 0.5955$$

And the same for CDF(16):

$$CDF(16) = CDF(17) - w_1PDF_1(17) + w_2PDF_2(17) =$$

$$= 0.5955 - 0.25\frac{e^{-10}10^{17}}{17!} - 0.75\frac{e^{-18}18^{17}}{17!} = 0.522$$

CDF(15):

$$CDF(15) = CDF(16) - w_1PDF_1(16) + w_2PDF_2(16) =$$

$$= 0.522 - 0.25\frac{e^{-10}10^{16}}{16!} - 0.75\frac{e^{-18}18^{16}}{16!} = 0.45$$

We got the first time with CDF>50% is 16. The time is 10:16.

Question 4 (25 pts)

A.

1. First note that E(RS(A))=307.5 and E(RS(B))=682.5. Also:

$$\sigma\big(RS(A)\big) = \sqrt{\frac{15 * 25 * 41}{12}} = 35.79$$

$$\sigma\big(RS(B)\big) = \sqrt{\frac{15 * 75 * 91}{12}} = 92.36$$

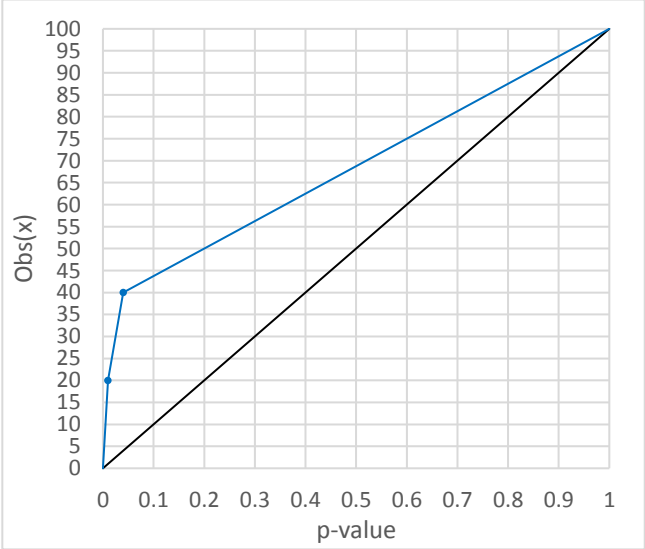
Now compute:

$$\mathrm{Z}\big(\mathrm{RS(A)}\big) = \frac{160 - 307.5}{35.79} = -4.12$$

$$\mathrm{Z}\big(\mathrm{RS(B)}\big) = \frac{200 - 682.5}{92.36} = -5.22$$

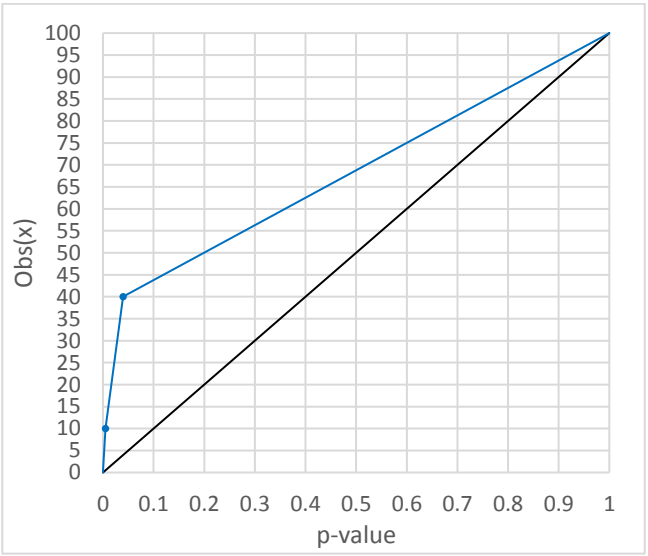
From which we conclude that the observation in Farm B is more significant.

2. From the conditions stated we conclude that the graph for Farm A must pass through the points (0.01, 20) for FDR=0.05 and (0.04, 40) for FDR=0.1. Other than this, we are free to draw ant monotone graph.
A possible graph can be:



From the conditions stated we conclude that the graph for Farm B must pass through the points (0.005, 10) for FDR=0.05 and (0.04, 40) for FDR=0.1. Other than this, we are free to draw ant monotone graph.

A possible graph can be:



B.

1. The distribution of X is:

	0	1	2	3	4
X	$\left(\frac{1}{2}\right)^4$	$4\left(\frac{1}{2}\right)^4$	$6\left(\frac{1}{2}\right)^4$	$4\left(\frac{1}{2}\right)^4$	$\left(\frac{1}{2}\right)^4$

$$H(X) = -2\left(\frac{1}{2}\right)^4 \log\left(\frac{1}{2}\right)^4 - 8\left(\frac{1}{2}\right)^4 \log\left(4\left(\frac{1}{2}\right)^4\right) - 6\left(\frac{1}{2}\right)^4 \log\left(6\left(\frac{1}{2}\right)^4\right)$$
$$= \frac{1}{2} + 1 + \frac{3}{2} - \frac{3}{8}\log 6 = 3 - \frac{3}{8}\log 6$$

2. Y attains 16 possible values. For such a random variable the maximal possible entropy is $\log 16 = 4$ and we get that for the uniform distribution. As the Binomial distribution is not uniform the statement is True.