Statistics and data analysis 2022
Final Exam (Bet)

Guidelines

- The exam will take place on Tuesday, 22 Feb 2022, at 15:45
- The exam will be held on campus.
- The total time of the exam is 1.5 hours (90 minutes).
- There are **3** (**THREE**) questions in the exam. You need to answer **2 (TWO)** of them.
- You can respond in English and/or Hebrew.
- 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 handheld calculators.
- No other auxiliary material can be used during the exam.
- Use normal approximation when appropriate and needed.

Good luck!

Question 1 (50 pts)

In this question $N(\mu, \sigma^2)$ stands for a normal distribution with mean μ and standard deviation σ.

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

Fast Fred takes time which is distributed as $N(30,25)$ minutes to repair a telephone line failure at a customer's home.

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

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

$X \sim N(30,25), Y \sim N(35,49), W \sim N(40,100)$

A. (10 points) Fred is due to arrive to repair your phone at 10AM tomorrow. How confident can you be that you will be done by 10:37?

$$P(X \leq 37) = P\left(Z \leq \frac{37-30}{5}\right) = \Phi\left(\frac{7}{5}\right) = 0.919$$

B. 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. (5 points) What is the distribution of the duration of repair in North Randomistan? Specify the density function.

$$f(t) = 0.4f_X(t) + 0.3f_Y(t) + 0.3f_W(t)$$
$$= \frac{0.4}{5\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-30}{5}\right)^2} + \frac{0.3}{7\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-35}{7}\right)^2} + \frac{0.3}{10\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-40}{10}\right)^2}$$

2. (5 points) 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.

$$F(t) = 0.4F_X(t) + 0.3F_Y(t) + 0.3F_W(t) = 0.4\Phi\left(\frac{t-30}{5}\right) + 0.3\Phi\left(\frac{t-35}{7}\right) + 0.3\Phi\left(\frac{t-40}{10}\right)$$

3. (15 points) If the repair starts at 10AM, which of the following times is the earliest time by which the customer can assume, with a 78% 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

10:42

$$F(23) = 0.4\Phi\left(-\frac{7}{5}\right) + 0.3\Phi\left(-\frac{12}{7}\right) + 0.3\Phi\left(-\frac{17}{10}\right) = 0.06 < 0.78$$

$$F(36) = 0.4\Phi\left(\frac{6}{5}\right) + 0.3\Phi\left(\frac{1}{7}\right) + 0.3\Phi\left(-\frac{4}{10}\right) = 0.62 < 0.78$$

$$F(42) = 0.4\Phi\left(\frac{12}{5}\right) + 0.3\Phi(1) + 0.3\Phi\left(\frac{2}{10}\right) = 0.82 > 0.78$$

C. (15 points) A new repair technician, Newton, started work in Randobezeq. Newton takes time which is distributed as $N(20,100)$ minutes to repair a telephone line failure at a customer's home.
For Fred and Newton, consider the time at which they will finish a repair with 97.725% certainty. For which of them is this time shorter?

The same time.
0.97725 is 2 standard deviations above the mean.
For Fred this means $30 + 2 * 5 = 40$
For Newton this means $20 + 2 * 10 = 40$

Question 2 (50 pts)

A. (20 points)

1. (10 points)

Let $\rho_S(u, v)$ be the Spearman correlation coefficient of the vectors u and v and $\rho_P(u, v)$ be the Pearson correlation coefficient of the vectors u and v.
Consider any 5 points $(x_1, y_1), \dots, (x_5, y_5)$
TRUE or FALSE:
If:

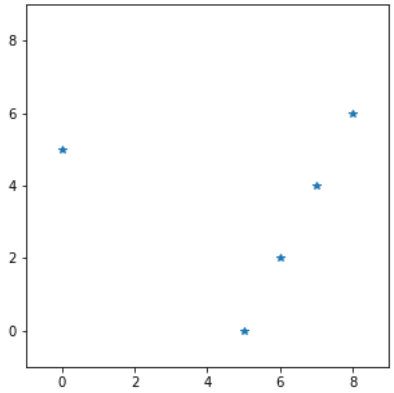
$$\rho_S\big((x_1, \dots, x_5), (y_1, \dots, y_5)\big) > 0$$

Then:

$$\rho_P\big((x_1, \dots, x_5), (y_1, \dots, y_5)\big) > 0$$

Prove your answer.

False:
In the following example, $\rho_S > 0, \rho_P < 0$



$$(x, y) = (0,5), (5,0), (6,2), (7,4), (8,6)$$
$$(R_x, R_y) = (1,4), (2,1), (3,2), (4,3), (5,5)$$
$$\rho_S = 1 - \frac{6 \cdot (9 + 1 + 1 + 1)}{5(25 - 1)} = 0.4$$
$$\bar{X} = 5.2, \bar{Y} = 3.4$$

$$\rho_P = \frac{(-5.2)(1.6) + (-0.2)(-3.4) + (0.8)(-1.4) + (1.8)(0.6) + (2.8)(2.6)}{\sqrt{(5.2^2) + (0.2^2) + (0.8^2) + (1.8^2) + (2.8^2)}\sqrt{(1.6^2) + (3.4^2) + (1.4^2) + (0.6^2) + (2.6^2)}}$$
$$= -\frac{0.4}{\sqrt{38.8}\sqrt{23.2}} = -0.13$$

2. (10 points)

Let $\tau(v, u)$ be the Kendall correlation on the vectors v and u.
Consider a set of $3n$ points:

$(x_1, y_1), \dots, (x_{3n}, y_{3n})$, where $n \geq 4$

Further assume that $x_i \neq x_j$ and $y_i \neq y_j \ \ \forall i \neq j$.
TRUE or FALSE:
If:

$\tau\big((x_1, \dots, x_n), (y_1, \dots, y_n)\big) = 0$ and
$\tau\big((x_{n+1}, \dots, x_{2n}), (y_{n+1}, \dots, y_{2n})\big) = 0$ and
$\tau\big((x_{2n+1}, \dots, x_{3n}), (y_{2n+1}, \dots, y_{3n})\big) = 0$

Then:

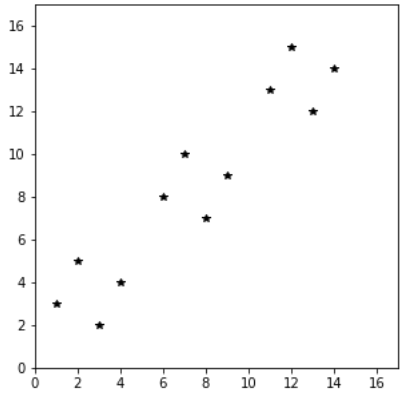
$\tau\big((x_1, \dots, x_{3n}), (y_1, \dots, y_{3n})\big) = 0$

Prove your answer.

False:

In the following example

$\tau\big((x_1, \dots, x_n), (y_1, \dots, y_n)\big) = 0,$
$\tau\big((x_{n+1}, \dots, x_{2n}), (y_{n+1}, \dots, y_{2n})\big) = 0,$
$\tau\big((x_{2n+1}, \dots, x_{3n}), (y_{2n+1}, \dots, y_{3n})\big) = 0$
and
$\tau\big((x_1, \dots, x_{3n}), (y_1, \dots, y_{3n})\big) > 0$



B. (30 points)
Consider three **independent** random variables, X, Y and W with the following distributions:

$$P(X = x) = \frac{1}{5}, \qquad x \in \{0,1,2,3,4\}$$
$$P(Y = y) = \frac{1}{3}, \qquad y \in \{0,5,10\}$$
$$P(W = w) = \frac{1}{2}, \qquad w \in \{0,15\}$$

1. (10 points)
Consider $Z = X + Y + W$. What is the distribution of Z?

$P(Z = z) = \frac{1}{30}, \qquad z \in \{0, \dots, 29\}$

2. Consider the 2D random variable (X, Y).
   a. (5 points) How many values does it attain?

15 values
$(x, y) \in \{0,1,2,3,4\} \times \{0,5,10\}$

   b. (5 points) What is $H(X, Y)$?

$$P\big((X,Y)=(x,y)\big)=\frac{1}{15},\qquad (x,y)\in\{0,1,2,3,4\}\times\{0,5,10\}$$

$$H(X,Y)=-\sum_{(x,y)}P(x,y)\log P(x,y)=-\sum_{i=1}^{15}\frac{1}{15}\log\left(\frac{1}{15}\right)=\log(15)=2.7$$

3. (10 points)
   Now assume that X,Y are not necessarily independent. Can $H(X,Y)$ be larger than the number you obtained in section 2b?

No. we showed in class that the entropy is maximal for the uniform distribution.

Question 3 (50 pts)

A. (16 points)
   Consider an experiment in which the p-values of 120 observations were calculated. Give an example for $p1,\dots,p120$ such that the experiment support reporting 15 results with FDR<0.05.

$p_1=p_2=\cdots=p_{15}=10^{-30}$
$p_{16}=p_{17}=\cdots=p_{120}=1-10^{-30}$
$FDR(15)=\frac{120*10^{-30}}{15}<0.05$
$FDR(16)=\frac{120*(1-10^{-30})}{16}>0.05$

B. (10 points) Consider the following experiment:
   a) Toss a fair coin 100 times.
   b) Compute the (one sided **left**) p-value of x, the observed number of 1s, under the fair coin null model. That is, compute $\mathrm{P}(X\le x)$, where $X\sim Binom(0.5,100)$.

   Repeat the experiment 20 independent times and sort the observed p-values:

$$p(1)\le p(2)\le\cdots\le p(20)$$

   What is $P(p(3)\le 0.1)$?
   Explain your answer.

$p_i\sim U(0,1),\qquad i=1,\dots,20$
$P(p_i\le 0.1)=0.1$
$P(p(3)\le 0.1)=P(for\ at\ least\ 3\ of\ p_1,\dots,p_{20}\ we\ have\ p_i\le 0.1)$
Let $X\sim Bin(0.1,20)$, so $P(p(3)\le 0.1)=P(X\ge 3)=0.32$

C. (24 points)
   1. (8 points) Match the 2 code segments to the 2 resulting figures. Explain your answer.
      Code segment 1:

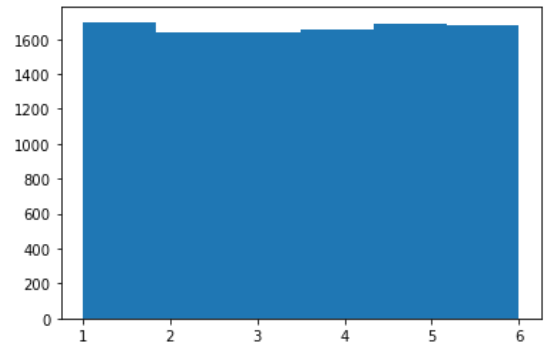
```
N = 10000
X = np.random.randint(1, 7, N)
plt.hist(X, bins=6)
```

      Code segment 2:

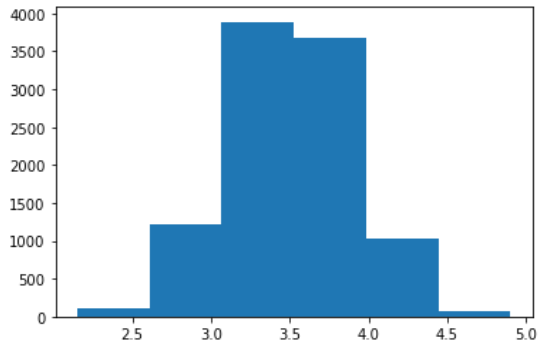
```
N = 10000
```

K = 20
Y = np.random.randint(1, 7, (N, K))
X = Y.mean(axis=1)
plt.hist(X, bins=6)

A



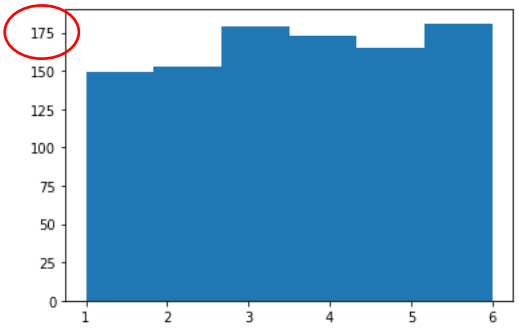
B



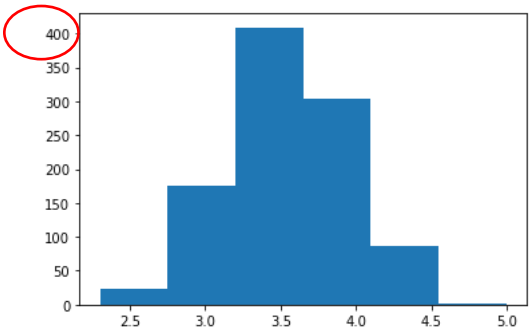
A = code segment 1 – histogram of a uniform distribution
B = code segment 2 – histogram of a normal distribution attained from the CLT for the mean.

2. (8 points) Now change *N* to be 1000 for both code segments. Schematically draw the resulting plots. Indicate relevant values on both axes. Explain.

A



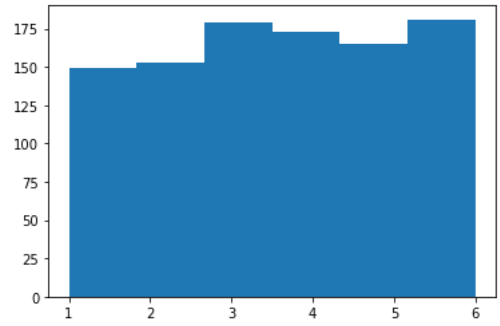
B



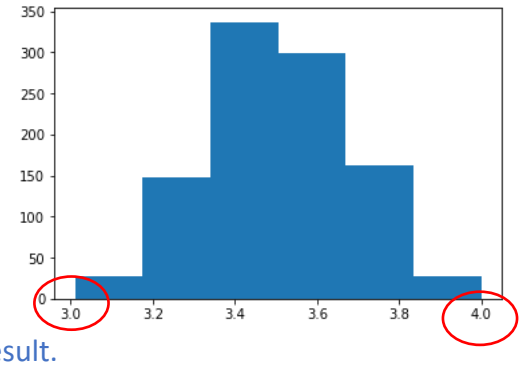
Same, only with less observations in each histogram

3. (8 points) Now keep *N* at 1000 and change *K* to be 100. Schematically draw the resulting plots. Indicate relevant values on both axes. Explain.

A



B



Same only with a narrower distribution in the CLT result.