Exam Integrity Statement – Online Exams

Course: _____ **Exam Date:** _____

Dear student,

In light of the Ministry of Health's guidelines and under the current circumstances, the Interdisciplinary Center Herzliya is doing everything in its power to maintain an academic routine, and accordingly this exam will be conducted online.

As you know, exam integrity is paramount in institutions of higher education and as such it is of great importance to IDC's administration, faculty and students; this, alongside striving for excellence and maintaining high professional and ethical academic quality.

The online exams are intended to reflect students' knowledge and proficiency in the course material and they must be conducted according to the required rules, fairly and abiding by the relevant instructions. Not abiding by the rules and instructions is a disciplinary offense that might also negatively affect all the other students who completed the assignment on their own merits.

The Interdisciplinary Center Herzliya will take any required action to ensure the integrity and fairness of the online exams is maintained.

Sincerely,
Interdisciplinary Center Herzliya

Student Statement

I hereby state and confirm the following, with regard to the exam noted above:

1. To the best of my knowledge, I am officially eligible to take this exam.
2. I hereby confirm that I am aware of the lecturer's instructions for the exam and that I completed the exam in compliance with these instructions.
3. I hereby confirm that I completed the exam alone, without consulting with or the assistance and cooperation of any other person.
4. I hereby confirm that I completed the exam within the required time framework, as noted in the lecturer's instructions; I submitted the exam at the end of the allotted time, and; I did not abuse the time extension awarded to students with this accommodation.
5. I am aware that Zoom software is operated during the exam as a supervisory method and I commit to not disrupt this action.
6. I am aware that in the efforts to maintain exam integrity I may be required by my lecturer, after receiving the exam and before receiving the grade, to explain some of my answers.

I am aware of the importance of conducting exams with integrity and fairness and according to the lecturer's instructions, and hereby confirm that I will comply with the above requirements. I am aware that non-compliance with these instructions and unfair conduct constitute a disciplinary offense at the IDC, with all that entails, including the possibility of cessation of studies.

I.D.: _____

Date: _____

Statistics and data analysis 2020
Final Exam (Bet)

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.
- 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.
- Use normal approximation when appropriate and needed.
- You can use hand held calculators and other resources as instructed.
- The total time of the exam is 3 (three) hours.
- Good luck!

Question 1 (25 pts)

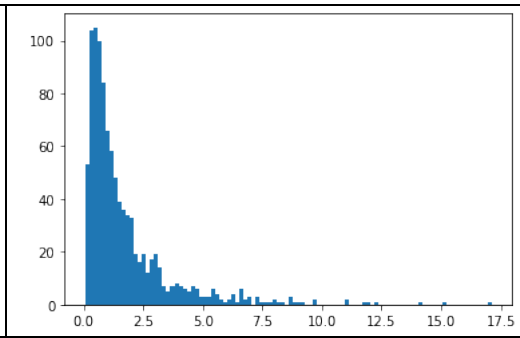
A. (6 pts)

Consider the following histograms came from a sample of size 1000 from some distribution. There are six of them.

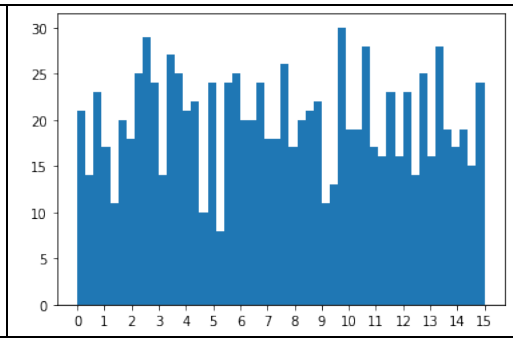
Determine a matching between the following distribution and the histograms: Normal, LogNormal, Poisson, Binomial, Gaussian Mixture, Uniform.

Indicate the matching clearly in your notebook. No calculations are needed.

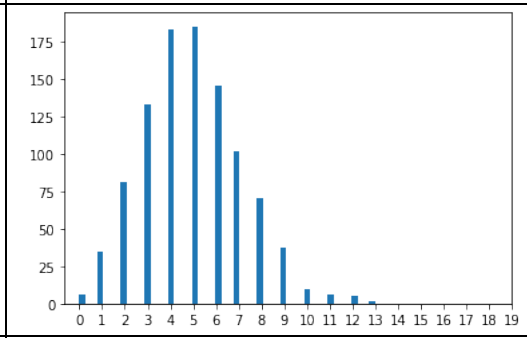
1



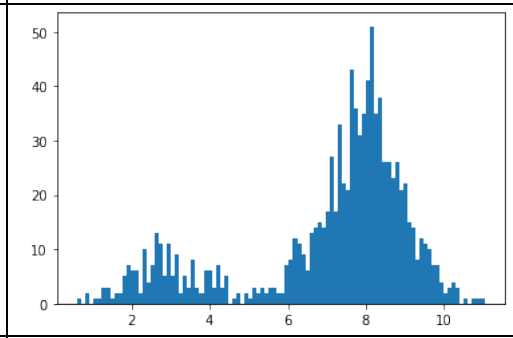
2



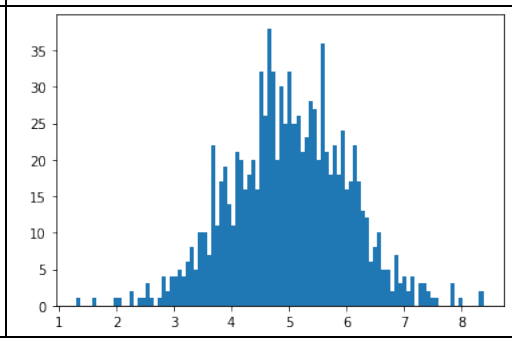
3



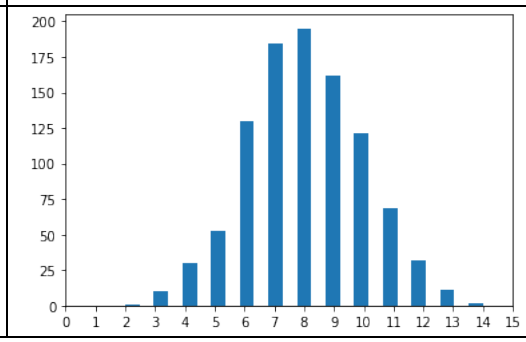
4



5



6



B. (11 pts)
Consider car safety scores coming from 3 factories A, B and C.
The data science team consider ONLY data from A and B.
Using WRS test they calculated that safety of A > safety of B with confidence of 0.95.
A week later, when the data of C also arrived, they calculated a WRS test on the data from ONLY B and C. They calculated that safety of B > safety of C with confidence of 0.95.

True or False
Safety of A > safety of C with confidence of 0.95 on the data from A and C ONLY.

Prove your answer.

C. (Total 8pts)
Similarly, consider safety data from two factories E and F.
For E we have 50 data points and for F we have 100 data points.

1. True or False
   $\forall \varepsilon > 0$ there exists data for which the safety of E is better than that of F with WRS p-value smaller than ε.
2. True or False
   $\forall \varepsilon > 0$ there exists data for which the safety of E is better than that of F with Student t-test p-value smaller than ε.

Prove your answers.

Question 2 (25 pts)

A. (Total 9 pts)
For each of the following, construct data as instructed or prove that this is impossible. In all cases assume that all the 18 values are different from each other.

$$u = u_1, \dots, u_9$$
$$v = v_1, \dots, v_9$$

1. (3 pts) $u_1, \dots, u_9$ and $v_1, \dots, v_9$ s.t. $\text{IQR}(u) = \text{IQR}(v)$ and $\text{std}(u) > \text{std}(v)$
2. (3 pts) $u_1, \dots, u_9$ and $v_1, \dots, v_9$ s.t. $\text{IQR}(u) = \text{IQR}(v)$ and $\text{std}(u) > \text{std}(v)$ and also $\max(u) < \max(v)$
3. (3 pts) $u_1, \dots, u_9$ and $v_1, \dots, v_9$ s.t. $\text{med}(u) = \text{med}(v) = 0$ and $\max(u) < \max(v)$ and $\max(u^2) > \max(v^2)$
   * In this section you can use once the same number in both vectors

B. (8 pts)
Let

$$X \sim Geo\left(\frac{1}{2}\right)$$
$$Y \sim Geo\left(\frac{1}{4}\right)$$

Assume that X and Y are independent.
Let

$$W = X + Y$$

Compute $P(W = 4)$

C. (8 pts) $X \sim NegBinom(r, p)$ where $0 < p < 1$.
Given that:

$$P(X = 2) = \frac{1}{16}$$
$$P(X = 3) = \frac{3}{32}$$

Compute the values of $E(X)$ and $V(X)$.

Question 3 (25 pts)

Let:

$$x = x_1, \dots, x_n$$
$$y = y_1, \dots, y_n$$

* All the numbers in each vector are different and $n \geq 5$.

In addition, let

$$z = z_1, \dots, z_n$$

Where

$$\forall i \;\; z_i = x_i + y_i$$

* $\tau(x, y)$ is the Kendal correlation between x and y

In all the following questions prove your answers.

A. (5 pts)
   True or False
   $$average(z) = average(x) + average(y)$$
B. (5 pts)
   True or False
   $$average(z^2) = average(x^2) + average(y^2)$$
C. (5 pts)
   True or False
   $$median(z) = median(x) + median(y)$$
D. (5 pts)
   True or False
   If
   $$\tau(x, y) = 1$$
   Then
   $$median(z) = median(x) + median(y)$$
E. (5 pts)
   True or False
   $$\exists x, y \;\; s.t \;\; \tau(x, y) = -1 \text{ and } median(z) = median(x) + median(y)$$

Question 4 (25 pts)

A research team preforms gene expression profiling to evaluate differences, in gene expression levels, comparing blood samples from people who are Covid-19 positive to blood samples from people who are Covid-19 negative.
They considered 10000 different genes and computed one sided WRS p-values for over expression in Covid-19.
These p-values are $0<p1<=p2<=\ldots<=p10000<=1$.
In all questions below assume: p5=0.00006, p10=0.00012, p21=0.0002, p500=0.02.

A. (5 pts) Is it possible that the team can report a set of features from this study with FDR better than 0.02? Explain your answer.
B. (5 pts) Is it possible that there is no set of features from this study with FDR better than 0.1? Explain your answer.
C. (5 pts) What is the best value they could possibly get for FDR(510)?
D. (5 pts) They observe FDR(200) = 0.01 ; What can you say about p200?
E. (5 pts) We know that they can't report 201 genes at FDR = 0.01. What can you say about p201?

Prove all your answers

Question 1 (25 pts)

A. (6 pts)
   1. LogNormal
   2. Uniform
   3. Poisson
   4. Gaussian Mixture
   5. Normal
   6. Binomial

B. (11 pts)
   False
   Consider the following data:
   $A = \{9, 10\}$
   $B = \{3,4,5,6,7,8\}$
   $C = \{1, 2\}$

   Calculating p-value of WRS(A,B)
   $p\_value = \frac{1}{\binom{8}{2}} = \frac{1}{28} < 0.05$
   Calculating p-value of WRS(B,C)
   $p\_value = \frac{1}{\binom{8}{2}} = \frac{1}{28} < 0.05$

   Now let's calculate the p-value of the WRS(A,C)
   $p\_value = \frac{1}{\binom{4}{2}} = \frac{1}{6} > 0.05$
   Q.E.D

C. (Total 8pts)
   1. False
      We'll get min p-value when the 50 data points of E will be smaller (or greater) than all the 100 data points of F.
      Let's now calculate this p-value:
      $$p\_value = \frac{1}{\binom{150}{50}}$$
      Now, consider $\varepsilon < \frac{1}{\binom{150}{50}}$.
      There is no arrangement of the data points that yields p-value lower than ε.
      Q.E.D
   2. True
      t-test statistics of two sets, is calculated from the means, stds and the number of samples. Given the number of samples in the question and assuming the stds are constant, we can shift the data of one of the sets (or both of them) as far apart as

we wish. Increasing the distance between the means will increase the t-test statistics and therefore decrease the p-value.
There is no limitation on this procedure and therefore $\forall \varepsilon > 0$ we can find a data for which the safety of E is better than that of F with Student t-test p-value smaller than ε.

Question 2 (25 pts)

A. (Total 9 pts)
For each of the following, construct data as instructed or prove that this is impossible. In all cases assume that all the 18 values are different from each other.

$$u = u_1, \dots, u_9$$
$$v = v_1, \dots, v_9$$

1. $u = (10,12,13,14,15,16,17,18,20)$
   $v = (1,2,3,4,5,6,7,8,9)$

2. $u = (0,2,3,4,5,6,7,8,10)$
   $v = (11,12,13,14,15,16,17,18,19)$

3. $u = (-15,-13,-12,-11,0,1,2,3,4)$
   $v = (-4,-3,-2,-1,0,11,12,13,14)$

B. (8 pts)
W is the convolution of X and Y.

$$P(W=4) = P(X=1 \cap Y=3) + (X=2 \cap Y=2) + (X=3 \cap Y=1)$$
$$P(X=1 \cap Y=3) = \frac{1}{2} * \left(\frac{3}{4}\right)^2 * \frac{1}{4} = \frac{9}{128}$$
$$P(X=2 \cap Y=2) = \left(\frac{1}{2}\right)^2 * \frac{3}{4} * \frac{1}{4} = \frac{3}{64}$$
$$P(X=1 \cap Y=3) = \left(\frac{1}{2}\right)^3 * \frac{1}{4} = \frac{1}{32}$$
$$P(W=4) = \frac{9}{128} + \frac{3}{64} + \frac{1}{32} = \frac{19}{128}$$

C. (8 pts)
r can be 1 or 2
If $r = 1$:

$$P(X=2) = (1-p)p = \frac{1}{16} \rightarrow p = 0.933 \; or \; 0.067$$
$$P(X=3) = (1-p)^2 p = 0.004 \; or \; 0.058 \neq \frac{3}{32}$$

Therefore, $r \neq 1$
$r = 2$:

$$P(X=2) = p^2 = \frac{1}{16} \rightarrow p = \frac{1}{4}$$
$$P(X=3) = 2(1-p)p^2 = 2 * \frac{3}{4} * \frac{1}{16} = \frac{3}{32}$$

$$E(X) = \frac{r}{p} = \frac{2}{\frac{1}{4}} = 8$$

$$V(X) = \frac{r(1-p)}{p^2} = \frac{2 * \frac{3}{4}}{\frac{1}{16}} = 24$$

Question 3 (25 pts)

A. (5 pts)
True

$$average(z) = average(x+y) = \frac{1}{n}\sum(x_i + y_i) = \frac{1}{n}\sum x_i + \frac{1}{n}\sum y_i$$
$$= average(x) + average(y)$$

B. (5 pts)
False

$$average(z^2) = average((x+y)^2) = average(x^2 + 2xy + y^2)$$
$$= \frac{1}{n}\sum x_i^2 + \frac{2}{n}\sum x_i y_i + \frac{1}{n}\sum y_i^2 = \frac{2}{n}\sum x_i y_i + average(x^2) + average(y^2)$$

If $\sum x_i y_i \neq 0$:

$$average(z^2) \neq average(x^2) + average(y^2)$$

C. (5 pts)
False

$x = (1,2,3,4,6)$
$y = (-1,-4,-2,-3,-5)$
$z = (0,-2,1,1,1)$

$median(x) = 3$
$median(y) = -3$
$median(z) = 1$

D. (5 pts)
True

WLOG assume $x_1 < x_2 < \cdots < x_n$.
Further, for simplicity let's assume that n is odd, $n = 2k+1$.
From $\tau(x,y) = 1$ we can conclude that $y_1 < y_2 < \cdots < y_n$.
Therefore, $median(x) = x_{k+1}$ and also $median(y) = y_{k+1}$.
We will now prove that $median(z) = x_{k+1} + y_{k+1}$.
For $1 \leq i \leq k$, we have $x_i < x_{k+1}$ and $y_i < y_{k+1}$.
For these $k = \frac{n-1}{2}$ indices we therefore have $z_i < x_{k+1} + y_{k+1}$.
Similarly, these are the only indices for which this holds.
Q.E.D

E. (5 pts)
True

$x = (1,2,3,4,5)$
$y = (-1,-2,-3,-4,-5)$
$z = (0,0,0,0,0)$

$median(x) = 3$
$median(y) = -3$
$median(z) = 0$

Question 4 (25 pts)

A.

Yes. For example, $p1 = 10^{-6}$:

$$FDR(1) = \frac{10^{-6} * 10^4}{1} = 10^{-2} < 0.02$$

B.

No. Let's look at p21:

$$FDR(21) = \frac{0.0002 * 10^4}{21} = 0.095 < 0.1$$

C.

$$p510 \geq p500 \rightarrow \min p510 = p500$$

$$FDR(510) \geq \frac{0.02 * 10^4}{510} = 0.39$$

D.

$$FDR(200) = \frac{p200 * 10^4}{200} = 0.01 \rightarrow p200 = 0.0002$$

E.

$$FDR(201) = \frac{p201 * 10^4}{201} > 0.01 \rightarrow p201 > 0.000201$$