Statistics and data analysis 2022
Final Exam (Alef)

Guidelines

- The exam will take place on Tuesday, 18 Jan 2022, at 15:45
- The exam will be online, via Zoom. You are required to connect to the Zoom meeting in the following link 15 minutes before the exam. https://idc-il.zoom.us/j/81574845874
- The total time of the exam is 1.5 hours (90 minutes) + 10 minutes for technical adjustments
- By the end of the exam time, you are required to submit a single PDF file to the course website.
- To avoid Moodle related technical issues, you can send your solution to following email address: idc.snda@gmail.com. This is **IN ADDITION** to the Moodle submission and will serve as a timestamp if technical difficulties arise. Note that only submissions done prior to the formal submission deadline will be considered.
- No other auxiliary material can be used during the exam.
- 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.
- Use normal approximation when appropriate and needed.
- You can use handheld calculators.
- Good luck!

Question 1 (50 pts)

<u>This question has 7 parts numbered A-G</u>

Let λ_1 be the **second** (unique, non-zero, indexed from left to right) digit of your ID number.
Let λ_2 be the **third** (unique, non-zero, indexed from left to right) digit of your ID number.
State your ID, $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$ clearly at the top of your solution

Let $X \sim \exp(\lambda_1)$ and $Y \sim \exp(\lambda_2)$ be two exponential RVs.
Remember that the density function of $A \sim \exp(\lambda)$ is given by $f(a) = \lambda e^{-\lambda a}, a \geq 0$

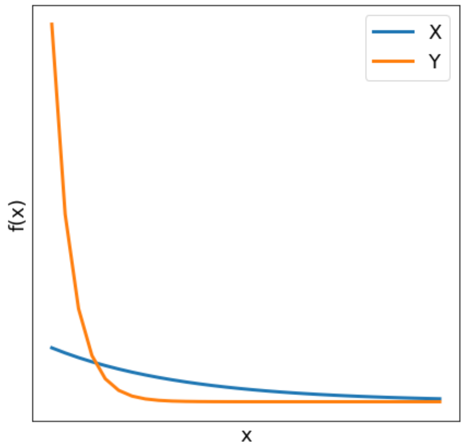
A. (8 pts) What is the expected value of X, $E(X)$?
   Show the complete derivation and give a value for the specific case of your ID number.
B. (7 pts) What is the median value of X, $med(X)$?
   Show the complete derivation and give a value for the specific case of your ID number.

Let $V = F(Y)$ where F is the CDF of Y.

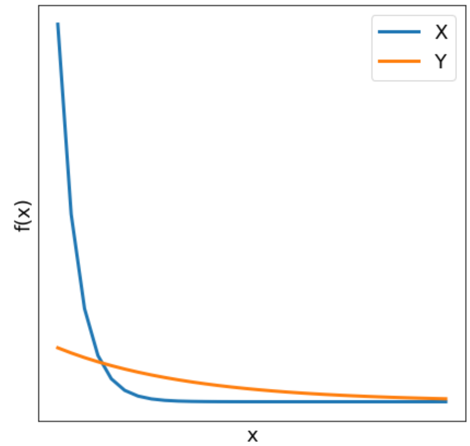
C. (7 pts) What is the range of V? (What values can V attain?)
D. (7 pts) What is the distribution of V?
   Show the formula for $P(V \leq v)$ for any value v. Show all calculations.
E. (7 pts) Assume that you have a function that generates random samples from V.
   Describe a method to generate random samples from Y.

F. (7 pts) Which of the following schematic density plots best represents the distributions of X and Y as defined using your ID number?
Explain your answer.

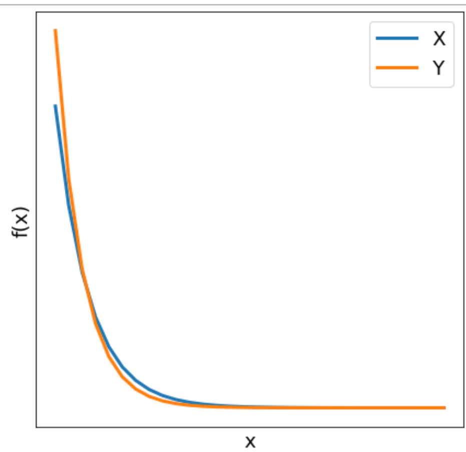
A



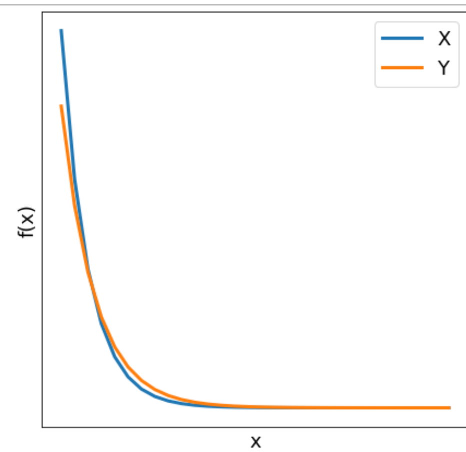
B



C



D



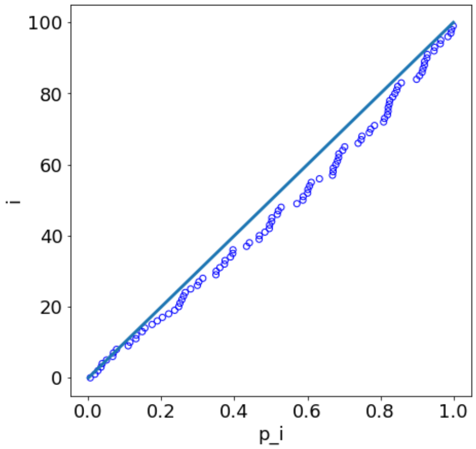
G. (7 pts) Consider the following pseudo code:

For $i = 1, \ldots, n$ *let* y_i *be a number drawn from the distribution of* Y
For $i = 1, \ldots, n$ *let* $q_i = P(X \leq y_i)$
Let $p_1, \ldots, p_n = \text{sort}(q_1, \ldots, q_n)$ *//meaning that* $p_1 \leq p_2 \leq \cdots \leq p_n$

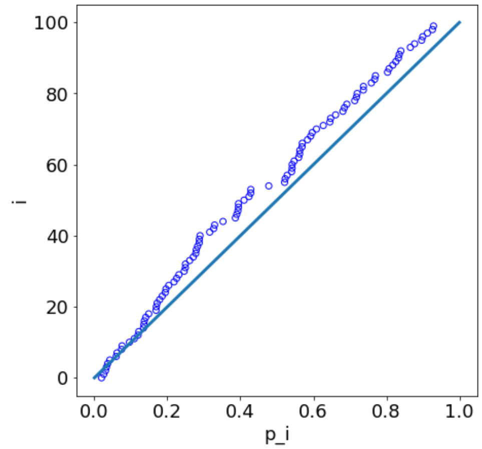
Which of the following plots best represents a scatter plot of $i = 1, \ldots, n$ against $p_1, \ldots, p_n$? Explain your answer.

* plt.scatter($[p_1, \ldots, p_n], [1, \ldots, n]$)

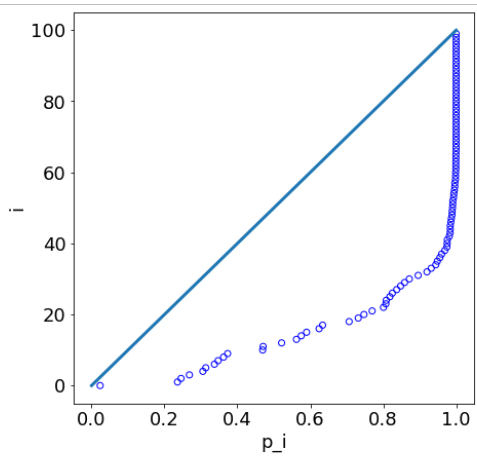
A



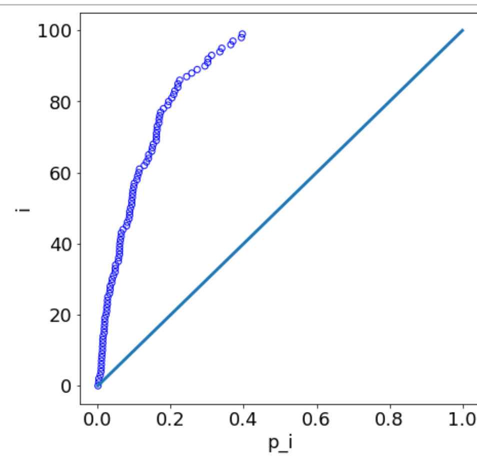
B



C



D



Question 2 (50 pts)

<u>This question has 4 parts numbered A-D</u>

A. (15 pts) Define a joint distribution over a pair of dice (X, Y) with 6 faces each that has the following properties:
   - The dice are NOT independent.
   - The marginals are uniform (fair)

   Show all calculations.

B. (9 pts) For the above RVs compute:
   1. $Cov(X, Y)$.
   2. $E(X + Y)$.
   3. The entropy $H(X + Y)$.

   Show all calculations.

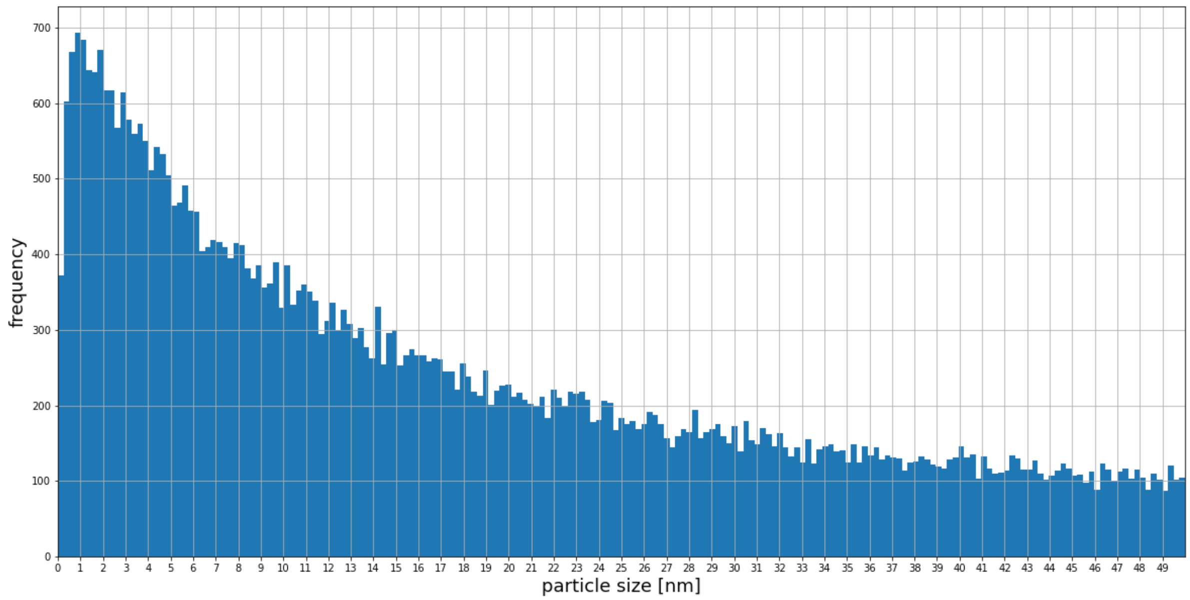
C. (13 pts) Let ID_3 be to the **third** (unique, non-zero, indexed from left to right) digit of your ID number.
   State your ID and $\mathbf{ID_3}$ clearly at the top of your solution.
   Let $N = 20 + ID_3$.
   Consider a vector of observed values $v = (v_1, \dots, v_N)$ where $v_1 < v_2 < \cdots < v_N$, coming from two classes C_1, C_2 and class associations vector: $L_1, \dots, L_N$ with matced ranking.
   Let $T(v)$ be the WRS statistic for the sum of ranks for C_1 obtained for v and $p(v)$ be the WRS **left side p-value** of v.
   Let $B = |C_1|$.
   What is the minimal and maximal values of B such that $\exists v;\ p(v) < 10^{-5}$?
   Explain your answer.

D. (13 pts) Let $X \sim NegBinom(r, p)$ where $0 < p < 1$.
   Given that:
   $P(X = 1) = 0$
   $P(X = 2) = \frac{1}{9}$
   Compute the values of $E(X)$ and $V(X)$.
   Show all calculations.

Question 3 (50 pts)

<u>This question has 5 parts numbered A-E</u>.

A scientist is generating nanoparticles for an experiment. She observes the following distribution of particle radii, in nm (nanometers):



This histogram representation of the distribution is calculated from 100,000 particles. The x-axis units are nm. The histogram is truncated at $50nm$. 51,503 particles of the 100,000 measured had radius ≥ 50 nm.

For the above data representing 100,000 particles, the scientist calculated empirical statistics. The empirical mean of the data is $\hat{\mu} = 403nm$ and the empirical 1st quartile $\hat{Q}_1 = 14nm$.

Upon looking at the histogram, the scientist decided to model the radii of the particles she generates using a random variable R with a lognormal distribution.

A. (10 pts) According to that model, what is the radius r so that $P(R \leq r) = 0.3$?

B. (10 pts) The experiment requires at least 40% of particles to have a radius smaller than $20nm$. Show, based on the above model, that the population generated here is therefore not adequate for the experiment.

C. (10 pts) The scientist can treat the particles and decrease all particle radii. A process that will reduce all particle radii by a factor of $\beta > 1$ ($R_{new} = \frac{1}{\beta}R$) will cost $\beta\ RCU$. How much will it cost to fulfill the experiment's requirement as stated above?

Show all your calculations.

Let $X \sim LogN(\mu_X, \sigma_X^2)$, $Y \sim LogN(\mu_Y, \sigma_Y^2)$ be two independent LogNormal random variables. Let $Z = XY$.

D. (10 pts) Express the CDF of Z in terms of Φ (the CDF of $N(0,1)$).

E. (10 pts) What is the PDF of Z?