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)

This question has 7 parts numbered A-G

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, λ_1 and λ_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.

$$E(X) = \int_0^\infty x\lambda_1 e^{-\lambda_1 x} =^{(*)} -xe^{-\lambda_1 x}\Big|_0^\infty - \int_0^\infty -e^{-\lambda_1 x}dx = 0 - \frac{1}{\lambda_1}e^{-\lambda_1 x}\Big|_0^\infty = \frac{1}{\lambda_1}$$

(*) integration by parts

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.

$m = med(X)$

$$F(x) = \int_0^x f(u)du = \int_0^x \lambda_1 e^{-\lambda_1 u}du = \left[-\frac{1}{\lambda_1}\lambda_1 e^{-\lambda_1 u}\right]_0^x = -e^{-\lambda_1 x} + 1 = 1 - e^{-\lambda_1 x}$$

$$P(X \leq m) = F(m) = 1 - e^{-\lambda_1 m} = 0.5 \Rightarrow e^{-\lambda_1 m} = 0.5 \Rightarrow m = -\frac{ln(0.5)}{\lambda_1} = \frac{ln\,2}{\lambda_1}$$

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?)

$V \in [0,1]$ since $v = F(y)$ is the probability $P(Y \leq y)$

D. (7 pts) What is the distribution of V?
   Show the formula for $P(V \leq v)$ for any value v. Show all calculations.

$$P(V \leq v) = P(F(Y) \leq v) = P\big(Y \leq F^{-1}(v)\big) = F\big(F^{-1}(v)\big) = v \Rightarrow v \sim Uniform([0,1])$$

The second equality follows from the positive monotonicity of F.
The others from the relevant definitions.

E. (7 pts) Assume that you have a function that generates random samples from V.
   Describe a method to generate random samples from Y.

Consider $T(v) = F^{-1}(v)$. Note that T is invertible and monotonically increasing.
Let $y \in \mathbb{R}$, $P(T(V) \leq y) = P\big(V \leq T^{-1}(y)\big) =^{(*)} T^{-1}(y) = F(y)$
(*) since $V \sim Uniform([0,1])$
Therefore $T(V)$ and Y have the same distribution.
Specifically, for our case and for $y, v \in \mathbb{R}$:

$$F(y) = 1 - e^{-\lambda_2 y}, \quad F^{-1}(v) = -\frac{\ln(1-v)}{\lambda_2}$$

To generate a sequence of points from Y:

1. Draw $v_1, \dots, v_N$ from V
2. Return $y_1 = F^{-1}(v_1), \dots, y_N = F^{-1}(v_N)$

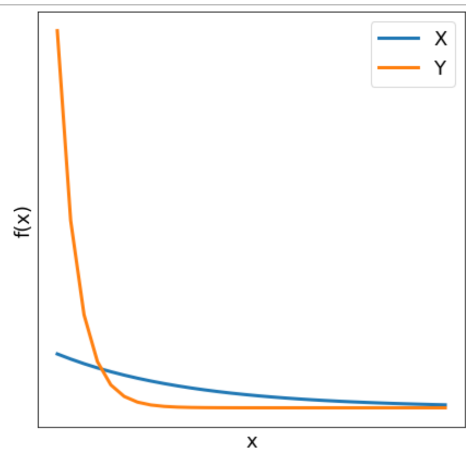
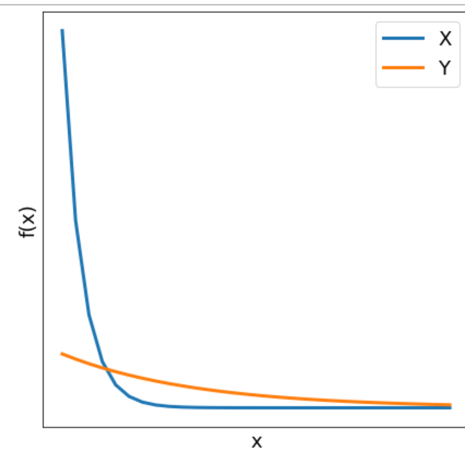
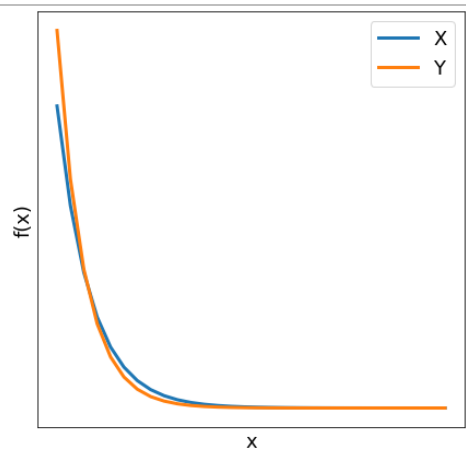
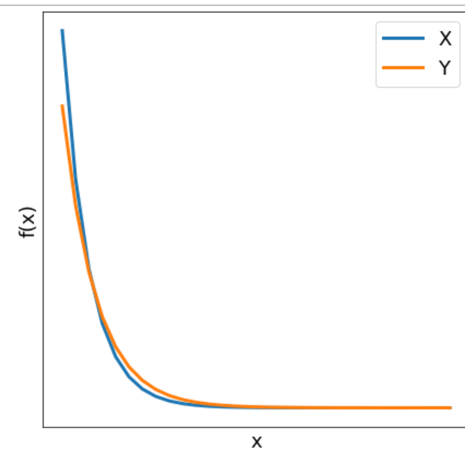
F. (7 pts) Which of the following schematic density plots best represents the distributions of X and Y as defined using your ID number?
Explain your answer.

For $\lambda_1 > \lambda_2$ we get $E(X) < E(Y)$ so the answer is either (B) or (D). Another explanation is that when $\lambda_1 > \lambda_2$ then the rate for X is higher and so most waiting times would be shorter putting most of the distribution in the lower values for X.

(B) if $\lambda_1 \gg \lambda_2$ and (D) if they are close in value.

Similarly, for $\lambda_1 < \lambda_2$ the answer is either (A) or (C).

(A) if $\lambda_1 \ll \lambda_2$ and (C) if they are close in value.

A	B
	
C	D
	

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

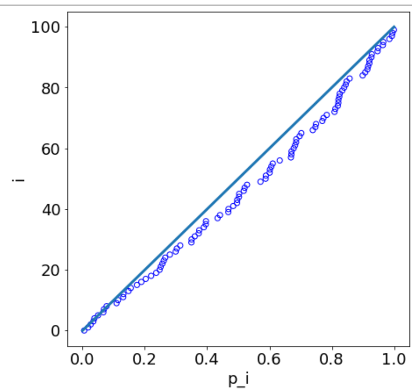
For $i = 1, \dots, n$ *let* y_i *be a number drawn from the distribution of* Y
For $i = 1, \dots, n$ *let* $q_i = P(X \leq y_i)$
Let $p_1, \dots, p_n = \text{sort}(q_1, \dots, q_n)$ *//meaning that* $p_1 \leq p_2 \leq \cdots \leq p_n$

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

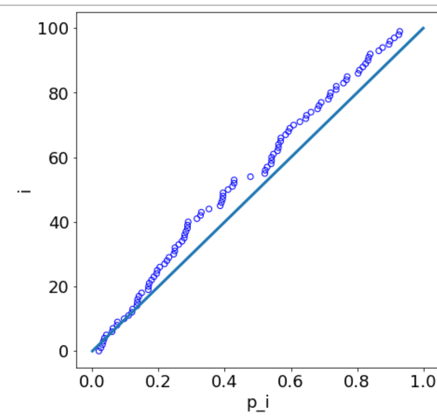
* plt.scatter($[p_1, \dots, p_n], [1, \dots, n]$)

This is similar to the experiment we showed in class. We draw data from Y and calculate one-sided left p-values while using X as the null model.
For $\lambda_1 > \lambda_2$, we expect to observe small values but the data contains larger values and so $P(X \leq y)$ is usually high. We therefore get underabundance of smaller p-values (below the line) so the answer is either (A) or (C). (C) if $\lambda_1 \gg \lambda_2$ and (A) if they are close in value.
For $\lambda_1 < \lambda_2$, we expect to observe large values but the data contains smaller values and so $P(X \leq y)$ is usually low. We therefore get overabundance of smaller p-values (above the line) so the answer is either (B) or (D). (D) if $\lambda_1 \ll \lambda_2$ and (B) if they are close in value.

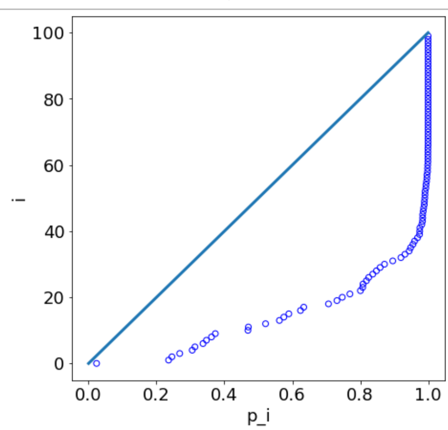
A



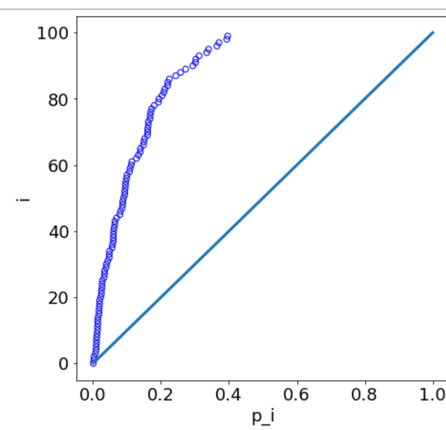
B



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.

Dice 2 / Dice 1	1	2	3	4	5	6	marginal
1	1/6	0	0	0	0	0	1/6
2	0	1/6	0	0	0	0	1/6
3	0	0	1/6	0	0	0	1/6
4	0	0	0	1/6	0	0	1/6
5	0	0	0	0	1/6	0	1/6
6	0	0	0	0	0	1/6	1/6
marginal	1/6	1/6	1/6	1/6	1/6	1/6	

The marginals are clearly uniform $P(X = i) = \sum_{j=1}^{6} P(X = i, Y = j) = \frac{1}{6} + 5 \cdot 0 = \frac{1}{6}, i = 1, \dots 6$ and similarly for $P(Y = j)$

X, Y are not independent, for example:

$$P(X = 1, Y = 1) = \frac{1}{6}$$
$$\neq$$
$$P(X = 1)P(Y = 1) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

* note that there are many other possible solutions.

B. (9 pts) For the above RVs compute:

1. $Cov(X, Y)$.

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$
$$= \frac{1}{6}(1 + 4 + 9 + 16 + 25 + 36) -$$
$$\frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) \cdot \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6)$$
$$= \frac{91}{6} - (3.5)(3.5) \approx 2.917$$

2. $E(X + Y)$.

   By linearity of expectation,

$$3.5 + 3.5 = 7$$

3. The entropy $H(X + Y)$.

Show all calculations.

Let $I = X + Y$. Note that I can only attain the values 2, 4, 6, 8, 10, 12, all with equal probability $\frac{1}{6}$ (all other values have probability 0).

$$H(I) = -\sum_{i=2}^{12} p(i)\log\big(p(i)\big)$$
$$(*)\ = 6\left(\frac{1}{6}\log(6)\right) = \log(6)$$

$(*)$ distributing the $(-)$ from outside the sum turns the $\log\left(\frac{1}{6}\right)$ into $\log(6)$

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 matched 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.
The minimal p-value is obtained when all B samples are at the top of the ordered vector and equals to p-value= $\frac{1}{\binom{20+ID_3}{B}}$.
We therefore need

$$10^{-5} > \frac{1}{\binom{20+ID_3}{B}}$$
$$10^{5} < \binom{20+ID_3}{B}$$

Such that $N = 20 + ID_3 \in [21,29]$, and B adheres to the requirement according to N.

Example solution: Let $ID_3 = 9$

$$100{,}000 = 10^5 < \binom{29}{B}$$

If we try $B = 4$, $\binom{29}{4} = 23{,}751 < 100{,}000$ so with $B = 4$ there is no vector v for which $p(v) < 10^{-5}$. If we try $B = 5$, $\binom{29}{5} = 118{,}755 > 100{,}000$ so with $B = 5$ there is at least one vector v for which $p(v) < 10^{-5}$.
Similarly (by symmetry of the choose function), $\binom{29}{25} < 100{,}000$ and $\binom{29}{24} > 100{,}000$
So for $B \in \{5,6,\dots,24\}$, $\exists v; p(v) < 10^{-5}$

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.

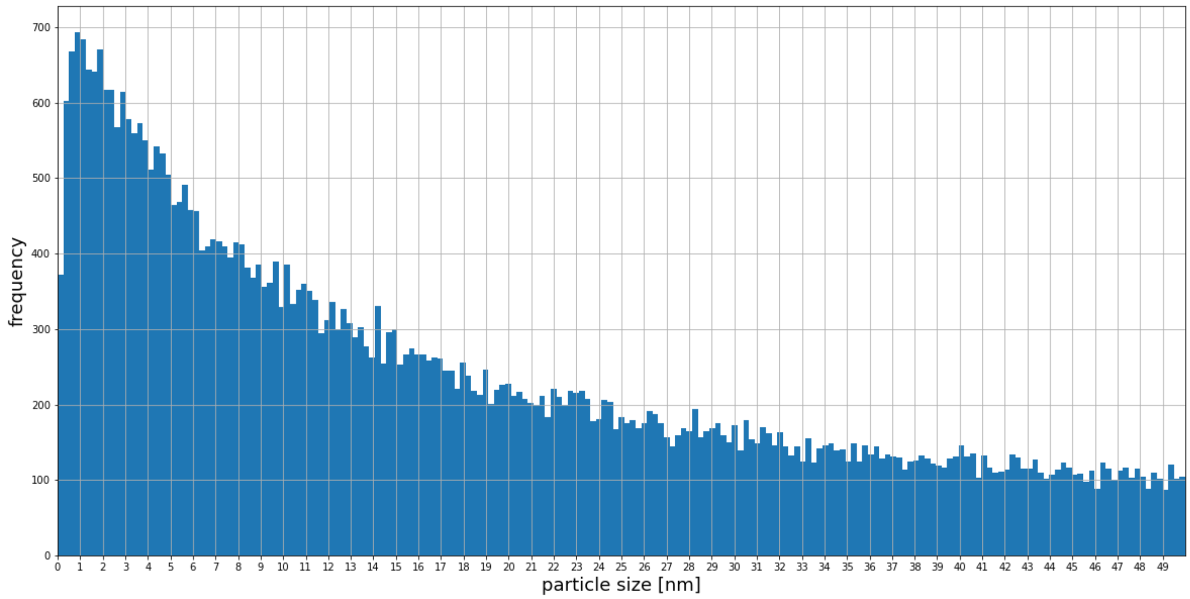
$P(X = 1) = 0$ tells us that $r \geq 2$
$P(X = 2) = \frac{1}{9}$ tells us that $r = 2$. since for $r \geq 3$ we have $P(X = 2) = 0$
Given $r = 2$ we get $P(X = 2) = p^2 = \frac{1}{9} \Rightarrow p = \frac{1}{3}$

$$E(X) = \frac{r}{p} = \frac{2}{\frac{1}{3}} = 6\,, \qquad V(X) = \frac{r(1-p)}{p^2} = \frac{2\left(1 - \frac{1}{3}\right)}{\frac{1}{3}^2} = \frac{\frac{4}{3}}{\frac{1}{9}} = 12$$

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 $P(R \leq r) = 0.3$?

We want to find r such that 30% of the distribution have radius less than r. In the standard normal distribution this corresponds to $\Phi^{-1}(0.3) = -0.52$ stds (using the z-score table).

Let μ and σ denote the mean and std of the underlying normal.

Two possible solutions:

A

From the plot we can observe that $Mode(R) = 1$:

$Mode(R) = 1 = e^{\mu - \sigma^2} \Rightarrow \mu = \sigma^2$

As $\hat{\mu} = 403 = e^{\mu + \frac{\sigma^2}{2}} \Rightarrow \ln(403) \approx 6 = \mu + \frac{\sigma^2}{2} = \mu + \frac{\mu}{2} \quad \Rightarrow \quad \mu = 4, \sigma = 2$

Therefore $r = e^{\mu+\Phi^{-1}(0.3)\cdot\sigma} = e^{4+(-0.52)\cdot 2} = e^{2.96} = 19.3nm$

B

We translate the information about the first quartile and about the part of the histogram to the right of 50 as follows

$$\frac{\ln 14 - \mu}{\sigma} = \Phi^{-1}(0.25) = -0.67$$

and

$$\frac{\ln 50 - \mu}{\sigma} = \Phi^{-1}(0.515) = 0.038$$

Solving for μ and σ we get $\mu = 3.9, \sigma = 1.8$ and a similar result for r.

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.

We require $P(R \leq 20) \geq 0.4$ which is equivalent to $P(X \leq \ln(20)) \geq 0.4$ where $X \sim N(4,2^2)$:

$$P(X \leq \ln(20)) = \Phi\left(\frac{\ln(20) - 4}{2}\right) \approx \Phi\left(-\frac{1}{2}\right) \approx 0.3 < 0.4$$

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.

The effect of the treatment is $R_{new} = \frac{1}{\beta}R = e^{\ln\left(\frac{1}{\beta}\right)} \cdot e^{X} = e^{X-\ln(\beta)}$ where $X \sim N(4,2^2)$

The new RV is lognormal with the underlying μ shifted by $\ln\left(\frac{1}{B}\right) = -\ln(\beta)$, with no change in the underlying σ.

Let β^* be the optimal β to meet the requirement.

Let R^* be the particle size distribution R_{new} obtained by using β^*. That is, $R^* = e^{X-\ln(\beta^*)}$.

We set $P(R^* \leq 20) = 0.4$ which is equivalent to $P(X - \ln(\beta^*) \leq \ln(20)) = 0.4$,

Which is equivalent to $\Phi\left(\frac{\ln(20)-4+\ln(\beta^*)}{2}\right) = 0.4$

$\Rightarrow \dfrac{\ln(20) - 4 + \ln(\beta^*)}{2} = \Phi^{-1}(0.4)$

$\Rightarrow \ln(\beta^*) = (-0.25) \cdot 2 + 4 - \ln(20)$

$\Rightarrow \ln(\beta^*) \approx 0.5$

$\Rightarrow \beta^* \approx 1.65$

$Cost = 1.65\ RCU$

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 $N(0,1)$).

Write: $X = e^U, Y = e^V\ where\ U \sim N(\mu_X, \sigma_X^2),\ \ and\ V \sim N(\mu_Y, \sigma_Y^2)$

Since X and Y are independent so are U and V (they are the result of taking a logarithm of the original pair). Therefore $U + V \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

Now write:

$$Z = XY = e^{U+V} \sim LogNormal(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Denote $\mu_Z = \mu_X + \mu_Y$ and $\sigma_Z = \sqrt{\sigma_X^2 + \sigma_Y^2}$

We then have $CDF_Z(z) = P(Z \le z) = P(e^W \le z) =^{(*)} P(W \le \ln(z))$

where $W \sim N(\mu_Z, \sigma_Z^2)$

(*) by monotonicity of the log function.

Finally, we have that:

$$CDF_Z = \Phi\left(\frac{\ln(z) - \mu_Z}{\sigma_Z}\right)$$

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

We obtain the PDF of Z by taking the derivative of the CDF:

$$PDF_Z(z) = \frac{1}{z\sigma_Z}\varphi\left(\frac{\ln(z) - \mu_Z}{\sigma_Z}\right)$$

Where φ is the standard normal density function.