University of Tehran,
College of Engineering

Homework 1

Introduction to Statistical Interference

Instructor: **Dr. Mohammadreza A. Dehaqani**

Erfan Panahi (*Student Number:* **810103084**)

Problem 1. Visa

Part a. The information regarding the applicants each year is shown in Table 1-1.

Year	All Applicants	Graduate Students
2021	308,613	121,000
2022	483,927	127,600
2023	780,884	314,000

Table 1-1. The information regarding the applicants each year

If we denote the number of annual applicants as A, among whom B are STEM graduates, the probability of winning each year is obtained as follows.

It is important to note that we must first calculate the probability of winning in the graduate group. If not successful in the graduate group, we subtract 20,000 from the total number of applicants, as 20,000 graduate students have already won the visa.

$$P_{winning\ in\ a\ year} = \frac{20000}{B} + \frac{B-20000}{B} \times \frac{65000}{A-20000}$$

Now, we calculate the probability of winning for the current year.

$$P_{2021} = \frac{20000}{121000} + \frac{101000}{121000} \times \frac{65000}{288613} \cong 0.3533$$

$$P_{2022} = (1-P_{2021})\left[\frac{20000}{127600} + \frac{107600}{127600} \times \frac{65000}{463927}\right] \cong 0.1778$$

$$P_{2023} = (1-P_{2021})(1-P_{2022})\left[\frac{20000}{314000} + \frac{294000}{314000} \times \frac{65000}{760884}\right] \cong 0.0764$$

Finally, we sum the three obtained probabilities together.

$$\boldsymbol{P(winning\ in\ 3\ years)} = P_{2021} + P_{2022} + P_{2023} \cong \mathbf{0.6075}$$

Part b. A non-STEM student only has one year to apply for the visa, so we use the probability calculated for a single year.

Using the 2021 data (as it’s their only year):

$$\boldsymbol{P(winning)} = P_{2021} \cong \mathbf{0.3533}$$

Since the probability is less than 50%, a non-STEM student **cannot reliably** depend on obtaining the H1B visa in one attempt.

Part c. The probability that, independent of participating in the H1B visa lottery, **marriage** would be the reason for obtaining the visa is as follows:

$$P_{marriage_{2021}} = 0.15$$

$$P_{marriage_{2022}} = \left(1 - P_{marriage_{2021}}\right) \times 0.15 = 0.85 \times 0.15 = 0.1275$$

$$P_{marriage_{2023}} = \left(1 - P_{marriage_{2021}}\right)\left(1 - P_{marriage_{2022}}\right) \times 0.15 = 0.7225 \times 0.15 \cong 0.1084$$

$$\boldsymbol{P(winning)_{marriage}} = P_{marriage_{2021}} + P_{marriage_{2022}} + P_{marriage_{2023}} \cong \mathbf{0.3859}$$

Now, considering the independence of the probabilities of marriage and winning the H1B visa lottery, we calculate **the probability of obtaining a green card within 3 years**.

$$\boldsymbol{P(getting\ \ a\ green\ card)} = \underbrace{P_{lottery}}_{0.5984} + \underbrace{P_{marriage}}_{0.3859} - \underbrace{P_{lottery \cap marriage}}_{\underbrace{P_{lottery} \times P_{marriage}}_{0.5984 \times 0.3859}} \cong \mathbf{0.7534}$$

Problem 2. Independence (Optional)

Part a. The probability mass function (PMF) of **Poisson** random variables N_1 and N_2 with parameters λ_1 and λ_2 are:

$$N_1 \sim Poisson(\lambda_1) \to P(N_1 = k) = e^{-\lambda_1}\frac{\lambda_1^k}{k!}, \qquad k > 0$$

$$N_2 \sim Poisson(\lambda_2) \to P(N_2 = k) = e^{-\lambda_2}\frac{\lambda_2^k}{k!}, \qquad k > 0$$

Given that N_1 and N_2 are independent, we can write their joint PMF as follows:

$$N_1 \sqcup N_2 \Longrightarrow P(N_1 = n, N_2 = k) = P(N_1 = k_1)P(N_2 = k_2)$$

Now we want to find the density function of the random variable $N = N_1 + N_2$.

$$P(N = n) = P(N_1 + N_2 = n) = \sum_{k=0}^{+\infty} P(N_1 = k, N_2 = n - k) = \sum_{k=0}^{+\infty} P(N_1 = k)P(N_2 = n - k)$$

$$= \sum_{k=0}^{+\infty} e^{-(\lambda_1+\lambda_2)}\frac{\lambda_1^k\lambda_2^{n-k}}{k!\,(n-k)!} = \frac{e^{-(\lambda_1+\lambda_2)}}{n!}\sum_{k=0}^{+\infty}\frac{n!}{k!\,(n-k)!}\lambda_1^k\lambda_2^{n-k}$$

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{n!}\sum_{k=0}^{+\infty}\binom{n}{k}\lambda_1^k\lambda_2^{n-k} = e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^n}{n!}$$

Finally, based on the obtained density function, we can determine the distribution of the random variable N.

$$P(N = n) = e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^n}{n!} \to \boldsymbol{N \sim Poisson(\lambda_1 + \lambda_2)}$$

Part b. The probability mass function (PMF) of **Exponential** random variables T_1 and T_2 with parameters λ_1 and λ_2 are:

$$T_1 \sim Exp(\lambda_1) \to f_{T_1}(t_1) = \lambda_1 e^{-\lambda_1 t_1}, \qquad t_1 > 0$$

$$T_1 \sim Exp(\lambda_2) \to f_{T_2}(t_2) = \lambda_2 e^{-\lambda_2 t_2}, \qquad t_2 > 0$$

To find the density of T, we use the *Jacobian method*. For this purpose, we define the auxiliary variables T and S as follows.

$$\begin{cases} T = g(T_1, T_2) = T_1 + T_2 \\ S = h(T_1, T_2) = T_2 \end{cases} \Longrightarrow \begin{cases} t_1 = t - s \\ t_2 = s \end{cases}$$

First, we calculate the Jacobian determinant.

$$J = \begin{vmatrix} \frac{\partial g}{\partial t_1} & \frac{\partial g}{\partial t_2} \\ \frac{\partial h}{\partial t_1} & \frac{\partial h}{\partial t_2} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1$$

Given that T_1 and T_2 are positive, we can say that $0 \le S \le T$.

Additionally, given the independence of T_1 and T_2, we know their joint density.

$$T_1 \sqcup T_2 \Longrightarrow f_{T_1T_2}(t_1,t_2) = f_{T_1}(t_1)f_{T_2}(t_2) = \lambda_1\lambda_2 e^{-\lambda_1 t_1 - \lambda_2 t_2}$$

$$f_{TS}(t,s) = \sum_i \frac{f_{T_1T_2}\left(t_{1_i}, t_{2_i}\right)}{|J|} = f_{T_1T_2}(t-s,s) = \lambda_1\lambda_2 e^{-\lambda_1(t-s)-\lambda_2 s} = \lambda_1\lambda_2 e^{-(\lambda_2-\lambda_1)s}e^{-\lambda_1 t}$$

$$f_{TS}(t,s) = \lambda_1\lambda_2 e^{-(\lambda_2-\lambda_1)s}e^{-\lambda_1 t}, \qquad 0 < s < t$$

$$f_T(t) = \int_{-\infty}^{+\infty} f_{TS}(t,s)ds = \int_0^t \lambda_1\lambda_2 e^{-(\lambda_2-\lambda_1)s}e^{-\lambda_1 t}ds = \lambda_1\lambda_2 e^{-\lambda_1 t}\int_0^t e^{-(\lambda_2-\lambda_1)s}ds$$

- if $\lambda_1 = \lambda_2 = \lambda$:

$$f_T(t) = \lambda^2 e^{-\lambda t}\int_0^t ds = \lambda^2 t e^{-\lambda t}$$

- if $\lambda_1 \neq \lambda_2$:

$$f_T(t) = \lambda_1\lambda_2 e^{-\lambda_1 t}\int_0^t e^{-(\lambda_2-\lambda_1)s}ds = \frac{\lambda_1\lambda_2 e^{-\lambda_1 t}}{(\lambda_2-\lambda_1)}\left(1 - e^{-(\lambda_2-\lambda_1)t}\right)$$
$$= \frac{\lambda_1\lambda_2}{(\lambda_2-\lambda_1)}\left(e^{-\lambda_1 t} - e^{-\lambda_2 t}\right)$$

Finally, we can write the density function of $T = T_1 + T_2$ as follows.

$$\boldsymbol{f_T(t) = \begin{cases} \dfrac{\lambda_1\lambda_2}{(\lambda_2-\lambda_1)}\left(e^{-\lambda_1 t} - e^{-\lambda_2 t}\right) & \text{if } \lambda_1 \neq \lambda_2 \\ \lambda^2 t e^{-\lambda t} & \text{if } \lambda_1 = \lambda_2 = \lambda \end{cases}, \quad t \geq 0}$$

Problem 3. Mean, Median, and Mode

Part a. According to Figure 3-1, we assume we have a distribution X with its center of symmetry at point c. Thus, we have:

$$f_X(c-x) = f_X(c+x)$$

$$F_X(c-x) = 1 - F_X(c+x)$$

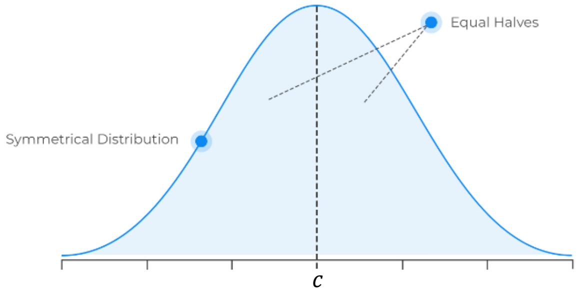


Figure 3-1. A Symmetrical Distribution ($f_X(x)$) – Unimodal

- **Mean:** Due to symmetry around c, contributions to the integral from $x < c$ and $x > c$ balance each other out, resulting in $\boldsymbol{E[X] = c}$. $(Mean = c)$
- **Median:** Since the distribution is symmetric about c, the cumulative distribution function (CDF) satisfies $F_X(c-x) = 1 - F_X(c+x)$. Therefore, c divides the distribution into two equal halves, making ccc the median as well. $(Median = c)$
- **Mode:** In a unimodal distribution with symmetry around c, the single peak (mode) of the distribution must occur at c, as any other location would violate the symmetry condition. Thus, c is also the mode of the distribution. $(Mode = c)$

Since c is the mean, median, and mode in a unimodal symmetric distribution, we conclude:

$$Mean = Median = Mode = c$$

Part b. **Multimodal Distribution:** A distribution with more than one mode. In other words, it has multiple peaks in its probability density function (PDF).

Figure 3-2 shows a bimodal distribution in which the mean, median, and modes are not the same.

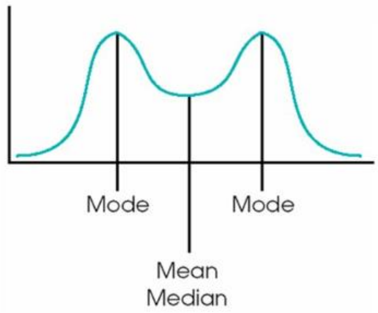


Figure 3-2. A Bimodal Symmetrical Distribution ($f_X(x)$) - (Graphical Example)

- **Mode:** In a multimodal distribution, there can be multiple modes (peaks). For instance, if a distribution has two modes, both will contribute to the mode measure. The **modes** could be at different points, and while the distribution is symmetric, these modes may not be exactly at the center of symmetry (mean and median).

- **Mean and Median:** In a **multimodal symmetric distribution**, the mean and median are still equal because the distribution is symmetric about a central point. However, the **mode** does not necessarily coincide with the mean or the median because there are multiple modes (peaks) in the distribution.

Another Example: Consider the **bimodal normal distribution**:

$$f(x) = 0.5 \times \mathcal{N}(-1, 1) + 0.5 \times \mathcal{N}(+1, 1)$$

where $\mu_1 = -1$ and $\mu_2 = 1$, and both modes have the same spread ($\sigma^2 = 1$). This is a symmetric bimodal distribution about $\mu = 0$, but there are two modes at $\mu_1 = -1$ and $\mu_2 = 1$, which are not at the same location.

- **Mean and Median:** $c = \mu = 0$
- **Modes:** $\mu_1 = -1$, $\mu_2 = 1$

Problem 4. A geometric point of view

Part a, b. If a point is chosen uniformly at random within the ellipse, the joint density function $f_{X,Y}(x,y)$ is constant over the area of the ellipse and zero outside. So, the joint density can be expressed as:

$$\boldsymbol{f_{XY}(x,y) = \begin{cases} \dfrac{1}{A_{ellipse}} = \dfrac{1}{\pi ab} & , \qquad \dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} \leq 1 \\ 0 & , \qquad o.w.(outside\ of\ the\ ellipse) \end{cases}}$$

To find the marginal density $f_X(x)$ of the x-coordinate, we integrate the joint density over the range of y:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy = \int_{-b\sqrt{1-\frac{x^2}{a^2}}}^{b\sqrt{1-\frac{x^2}{a^2}}} \frac{1}{\pi ab} dxdy = \frac{2\sqrt{1-\frac{x^2}{a^2}}}{\pi a} \rightarrow \boldsymbol{f_X(x) = \frac{2\sqrt{1-\frac{x^2}{a^2}}}{\pi a}} , \qquad \boldsymbol{|x| \leq a}$$

To find the marginal density $f_Y(y)$ of the y-coordinate, we integrate the joint density over the range of x:

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dx = \int_{-a\sqrt{1-\frac{y^2}{b^2}}}^{a\sqrt{1-\frac{y^2}{b^2}}} \frac{1}{\pi ab} dxdy = \frac{2\sqrt{1-\frac{y^2}{b^2}}}{\pi b} \rightarrow \boldsymbol{f_Y(y) = \frac{2\sqrt{1-\frac{y^2}{b^2}}}{\pi b}} , \qquad \boldsymbol{|y| \leq b}$$

Part c. If a point is chosen

$$P\left(|X| \leq \frac{a}{2}\right) = P\left(-\frac{a}{2} \leq X \leq \frac{a}{2}\right) = \int_{-\frac{a}{2}}^{\frac{a}{2}} \frac{2\sqrt{1-\frac{x^2}{a^2}}}{\pi a} dx$$

$$\begin{cases} x = a\sin(\theta) \\ dx = a\cos(\theta)\, d\theta \end{cases}$$

$$P\left(|X| \leq \frac{a}{2}\right) = \frac{2}{\pi}\int_{-\frac{\pi}{6}}^{\frac{\pi}{6}} \cos^2(\theta)\, d\theta = \frac{2}{\pi}\int_{-\frac{\pi}{6}}^{\frac{\pi}{6}} \left(\frac{1}{2} + \frac{1}{2}\cos(2\theta)\right) d\theta = \frac{1}{3} + \frac{1}{\pi}\int_{-\frac{\pi}{6}}^{\frac{\pi}{6}} \cos(2\theta)\, d\theta$$

$$= \frac{1}{3} + \frac{1}{2\pi}\sin(2\theta)\,\Big|_{-\frac{\pi}{6}}^{\frac{\pi}{6}} = \frac{1}{3} + \frac{\sqrt{3}}{2\pi}$$

$$\boldsymbol{P\left(|X| \leq \frac{a}{2}\right) = \frac{1}{3} + \frac{\sqrt{3}}{2\pi}}$$

Problem 5. Inequalities

Part a. Markov inequality

The **Markov inequality** states that for any non-negative random variable X and any $a > 0$:

$$P(X \geq a) \leq \frac{E[X]}{a}$$

1. Assuming that for the random variable X we have $E(X) = 5$, we want to find the probability $P(X \leq 10)$:

$$a = 10 \rightarrow P(X \geq 10) \leq \frac{5}{10} \rightarrow \boldsymbol{P(X \geq 10) \leq \frac{1}{2}}$$

2. The Markov inequality provides an upper bound rather than an exact probability because it only uses $E(X)$ to make the estimate. It does not take into account any other information about the distribution of X, such as its variance, skewness, or specific shape.

   The inequality is derived based on the assumption that all the probability mass of X could be concentrated at or above a, which is often not the case in practice. Therefore, the actual probability $P(X \geq a)$ could be much smaller than the upper bound provided by the Markov inequality.

3. For **random variables with low variance**, most of the probability mass is clustered close to the mean. In these cases, the probability of X being far from $E(X)$ is small, so the Markov bound is not tight.

Part b. Chebyshev inequality

The **Chebyshev inequality** states that for any random variable X with mean $E(X) = \mu$ and variance $var(X) = \sigma^2$, for any $k > 0$:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

1. Assuming that for the random variable X we have $E[X] = 20$ and $Var(X) = 36$, we want to find the probability $P(|X - 20| \geq 12)$:

$$\mu = 20\,, \sigma = 6\,, k = 2 \rightarrow P(|X - 20| \geq 12) \leq \frac{1}{2^2} \rightarrow \boldsymbol{P(|X - 20| \geq 12) \leq \frac{1}{4}}$$

2. n this calculation, k represents the number of standard deviations away from the mean. Specifically, $k = 2$ means that we are interested in the probability that X deviates by at least two standard deviations from the mean. The choice of k directly impacts the upper bound, as the bound is inversely proportional to k^2. Larger values of k (meaning we're looking at points further from the mean) will result in a smaller upper bound, while smaller values of k yield larger upper bounds.

3. If we were to draw the distribution of X, Chebyshev's inequality would likely overestimate the actual probability of $| X - 20 | \geq 12$. This is because Chebyshev's inequality is **general** and applies to any distribution with finite mean and variance. It does not assume anything about the shape of the distribution of X.

Part c. Cauchy-Schwarz inequality

To use the **Cauchy-Schwarz inequality** to find an upper bound for $E(XY)$, we start with the inequality in the context of expectations. The Cauchy-Schwarz inequality states that for any random variables X and Y:

$$(E[XY])^2 \leq E(X^2)E(Y^2)$$

1. Assuming that for the random variables X and Y we have $E[X] = 2, E[Y] = 3, E(X^2) = 8$, and $E(Y^2) = 12$, we want to find an upper bound for $E(XY)$:

$$\left(E(XY)\right)^2 \leq 8 \times 12 = 96 \rightarrow |E(XY)| \leq 4\sqrt{6}$$

2. The Cauchy-Schwarz bound only considers the magnitudes of X and Y and does not account for the **correlation** between X and Y. Here are situations where $E(XY)$ might be much lower than this upper bound:

   - If <u>X and Y are uncorrelated</u> $(cov(X,Y) = 0)$, then $E(XY) = E(X)E(Y) = 6$, which is significantly lower than $4\sqrt{6}$. In fact, the upper bound provided by Cauchy-Schwarz is achieved only when X and Y are perfectly correlated $(\rho_{XY} = \pm 1)$.
   - If X and Y are negatively correlated, $E(XY)$ could be even lower than 6, potentially approaching zero or even taking a negative value, depending on the strength of the negative correlation.

Problem 6. Sort, sort, sort (Simulation_Problems.ipynb file)

Part a. To implement this section, we need the *'sorting_algorithms.py'* file, which contains the sorting functions required by the problem. (These codes were found here.)

According to the problem's requirements, we generate the target list and measure the sorting time for each algorithm in each iteration. Figure 6-1 shows the distribution of sorting time for each algorithm along with their mean and median.

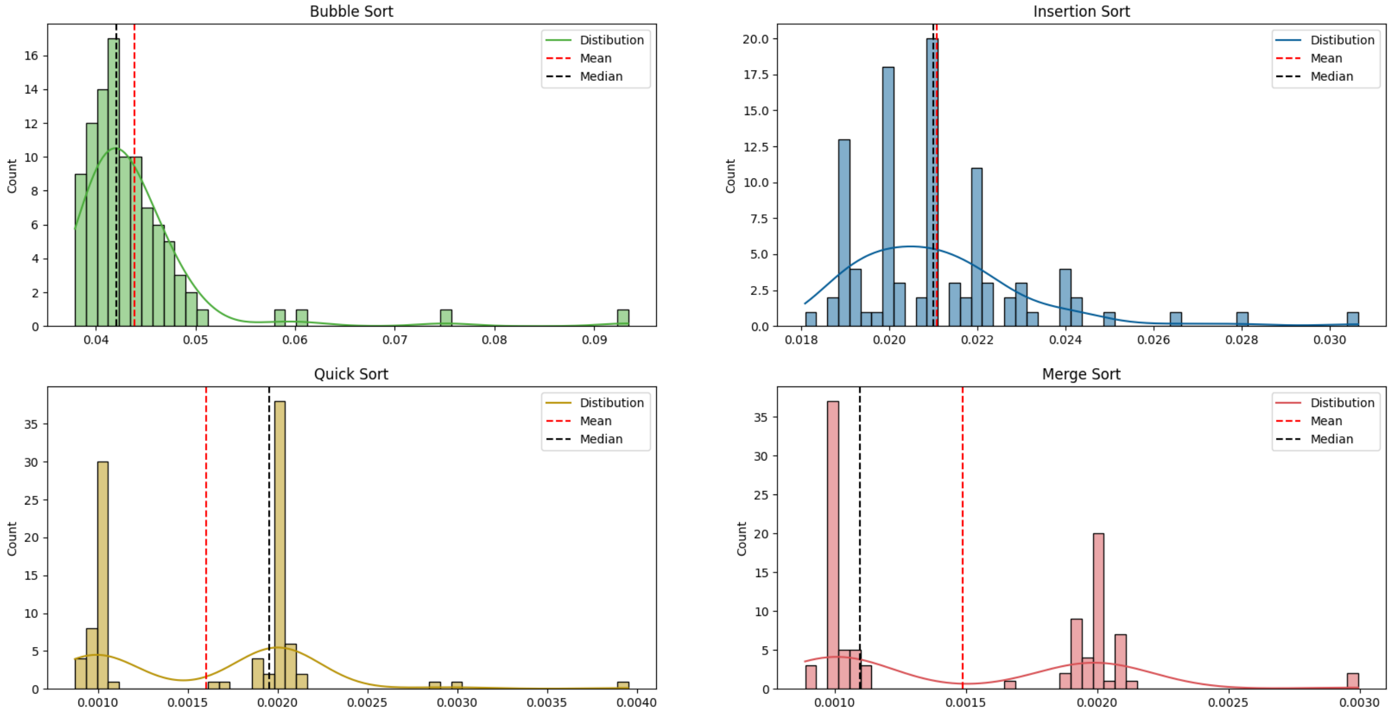
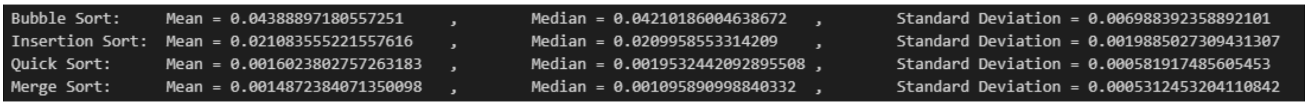


Figure 6-1. Distribution of sorting time for each algorithm along with their mean and median

Figure 6-2 presents the mean, median, and standard deviation of the sorting time for each algorithm.



```
Bubble Sort:     Mean = 0.04388897180557251    ,      Median = 0.04210186004638672    ,      Standard Deviation = 0.006988392358892101
Insertion Sort:  Mean = 0.021083555221557616   ,      Median = 0.0209958553314209     ,      Standard Deviation = 0.0019885027309431307
Quick Sort:      Mean = 0.0016023802757263183  ,      Median = 0.0019532442092895508 ,      Standard Deviation = 0.000581917485605453
Merge Sort:      Mean = 0.0014872384071350098  ,      Median = 0.001095890998840332   ,      Standard Deviation = 0.0005312453204110842
```

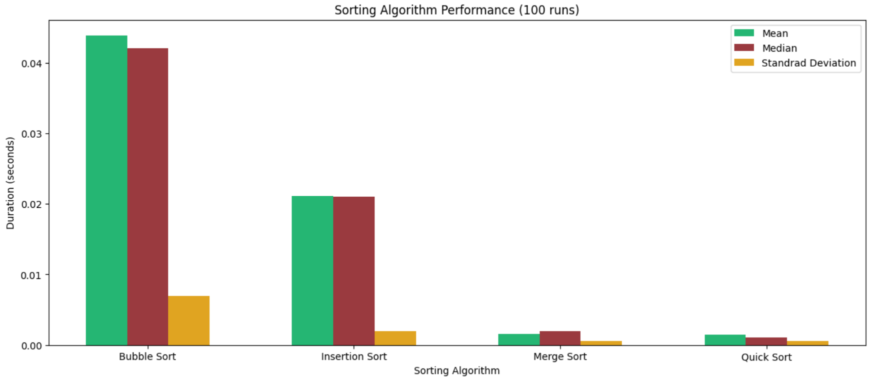


Figure 6-2. Mean, median, and standard deviation of the sorting time for each algorithm

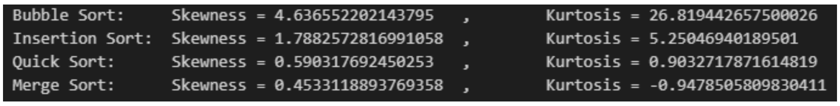
Based on the obtained results, we can say that the **Merge Sort** algorithm is the fastest in terms of average. Therefore, in terms of average speed, the algorithms are ranked as follows.

$$\textbf{Based on the mean duration:}\quad \text{Merge sort} < \text{Quick sort} < \text{Insertion sort} < \text{Bubble sort}$$

Regarding stability, as indicated by the **standard deviation**, the **Merge Sort** algorithm is the most stable, while the **Bubble Sort** algorithm is the most volatile.

$$\textbf{Based on the standard deviation:}\quad \underbrace{\text{Merge sort}}_{\textit{The most stable}} > \text{Quick sort} > \text{Insertion sort} > \underbrace{\text{Bubble sort}}_{\textit{The most volatile}}$$

Part b. Figure 6-3 presents the Kurtosis and Skewness of the sorting time for each algorithm.



```
Bubble Sort:     Skewness = 4.636552202143795   ,        Kurtosis = 26.819442657500026
Insertion Sort:  Skewness = 1.7882572816991058  ,        Kurtosis = 5.25046940189501
Quick Sort:      Skewness = 0.590317692450253   ,        Kurtosis = 0.9032717871614819
Merge Sort:      Skewness = 0.4533118893769358  ,        Kurtosis = -0.9478505809830411
```

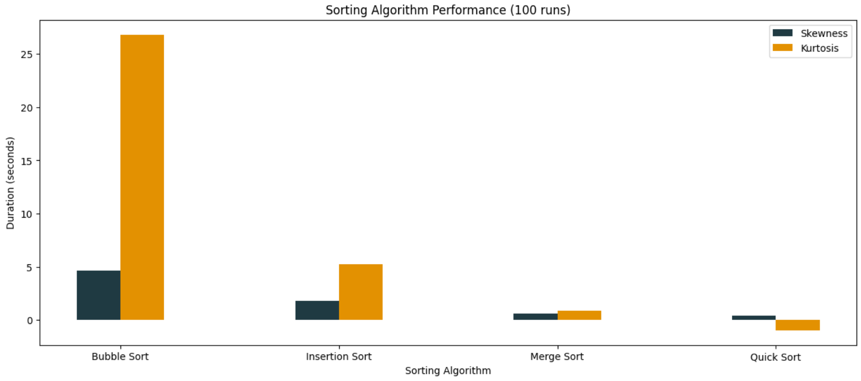


Figure 6-3. Kurtosis and Skewness of the sorting time for each algorithm

- **Skewness:** The distributions for **bubble sort** and **insertion sort** are often right-skewed because, in the worst-case scenario, they can take significantly longer to sort a list, resulting in a longer tail to the right.
- **Algorithm with Least Kurtosis:** **Merge sort** typically has the least kurtosis. This suggests that its distribution of sorting times is close to normal with fewer outliers and consistent performance. Low kurtosis means that the algorithm's duration does not exhibit many extreme deviations from the mean, reinforcing its reliability.

Part c. Figure 6-4 presents the boxplot for the sorting time for each algorithm.

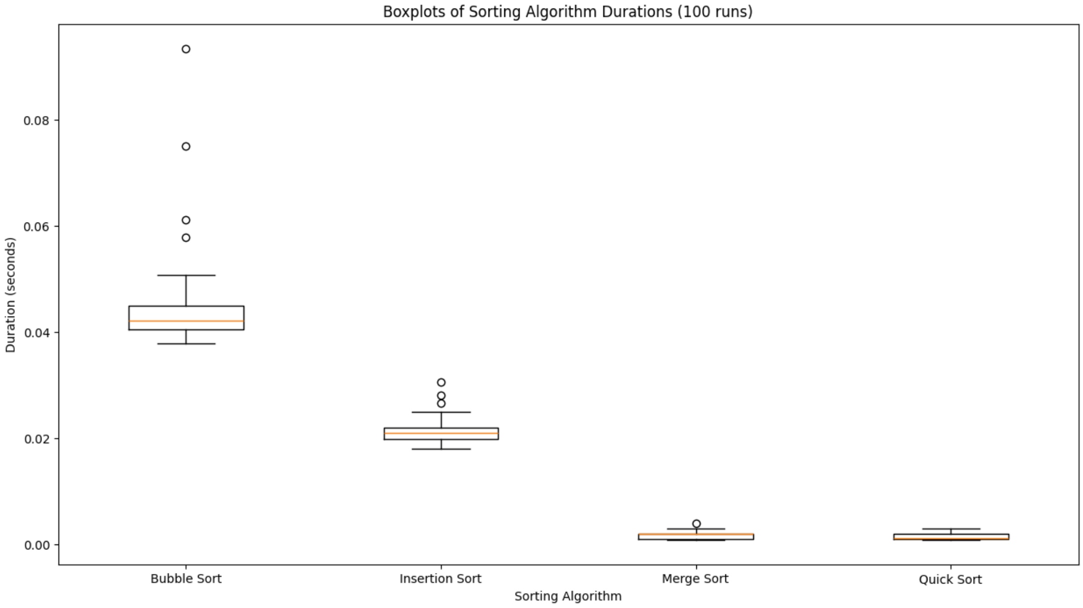


Figure 6-4. Boxplots for the durations of each algorithm

The distributions for **bubble sort** and **insertion sort** often have the most outliers because, in the worst-case scenario, they can take significantly longer to sort a list. This results in some execution times being much higher than the others, creating extreme values or outliers.

Problem 7. Cookie Monster (Simulation_Problems.ipynb file)

Part a. According to the given descriptions in the exercise, after reading the image using the *cv2* package, we add three types of noise "uniform, Gaussian, and exponential" to the image for six different noise ratios $(0.1, 0.15, 0.2, 0.3, 0.4, 0.5)$.

Figure 7-1 shows the noisy image for each noise type and ratio. As observed, as the noise ratio increases, less content of the image is visible.

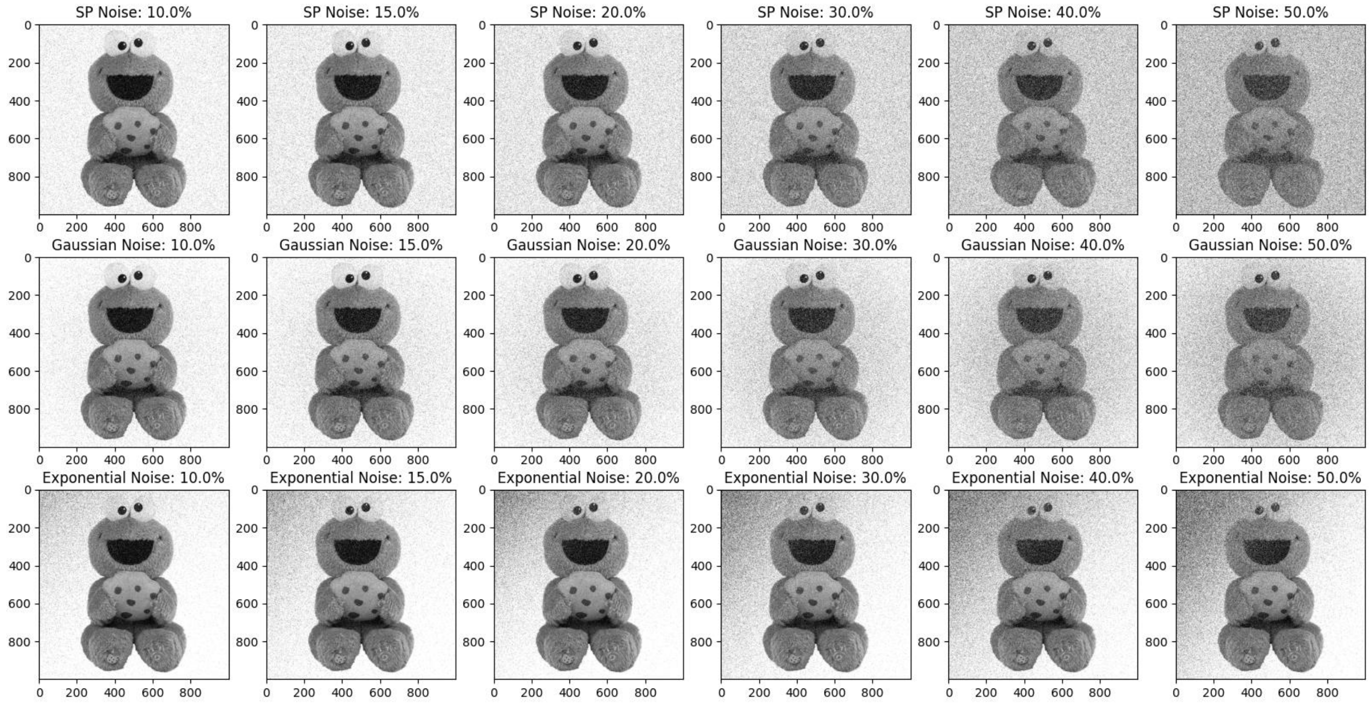


Figure 7-1. The noisy image for each noise type and ratio

Part b. In this section, we consider a square as shown in Figure 7-2, which includes the face of Cookie Monster.

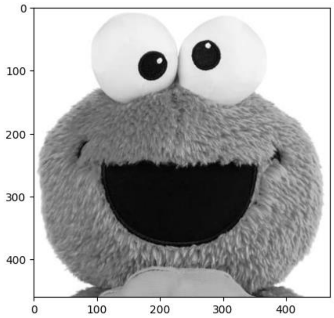
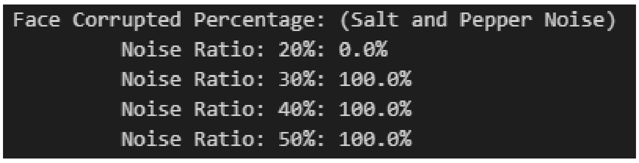


Figure 7-2. The face of the cookie monster

Now, for 100 iterations and for each noise ratio, we analyze how much of the face is affected by noise. If 20% or more of the face is noisy, we consider the image to be corrupted.

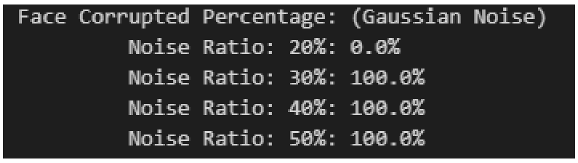
Figure 7-3 shows that with the addition of uniform noise, only at the 20% noise ratio is the image not corrupted.



```
Face Corrupted Percentage: (Salt and Pepper Noise)
        Noise Ratio: 20%: 0.0%
        Noise Ratio: 30%: 100.0%
        Noise Ratio: 40%: 100.0%
        Noise Ratio: 50%: 100.0%
```

Figure 7-3. Percentage of corrupted image (Uniform)

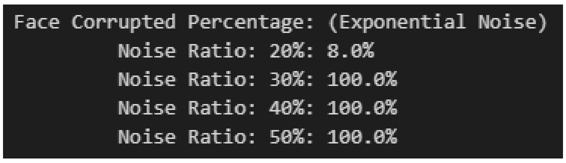
Figure 7-4 shows that with the addition of Gaussian noise, only at the 20% noise ratio is the image not blurred.



```
Face Corrupted Percentage: (Gaussian Noise)
        Noise Ratio: 20%: 0.0%
        Noise Ratio: 30%: 100.0%
        Noise Ratio: 40%: 100.0%
        Noise Ratio: 50%: 100.0%
```

Figure 7-4. Percentage of corrupted image (Gaussian)

Figure 7-5 shows that with the addition of exponential noise, only at the 20% noise ratio is the image not blurred.



```
Face Corrupted Percentage: (Exponential Noise)
        Noise Ratio: 20%: 8.0%
        Noise Ratio: 30%: 100.0%
        Noise Ratio: 40%: 100.0%
        Noise Ratio: 50%: 100.0%
```

Figure 7-5. Percentage of corrupted image (Exponential)

We conclude that for noise ratios higher than 20%, the image becomes corrupted, and as this ratio increases, less content is visible in the image. In the case of Gaussian noise, this depends on the mean and variance of the distribution, and in exponential distribution, it depends on scaling the distribution to the horizontal and vertical indices. Generally, it can be said that on average, all three noise models behave similarly and corrupt the image to a similar extent.

Part c. Finally, we intend to restore the image by adding a median filter. For this purpose, we use the *scipy.ndimage* package. We filter the images for the noise ratios from part b. Figure 7-6 shows the filtered images.

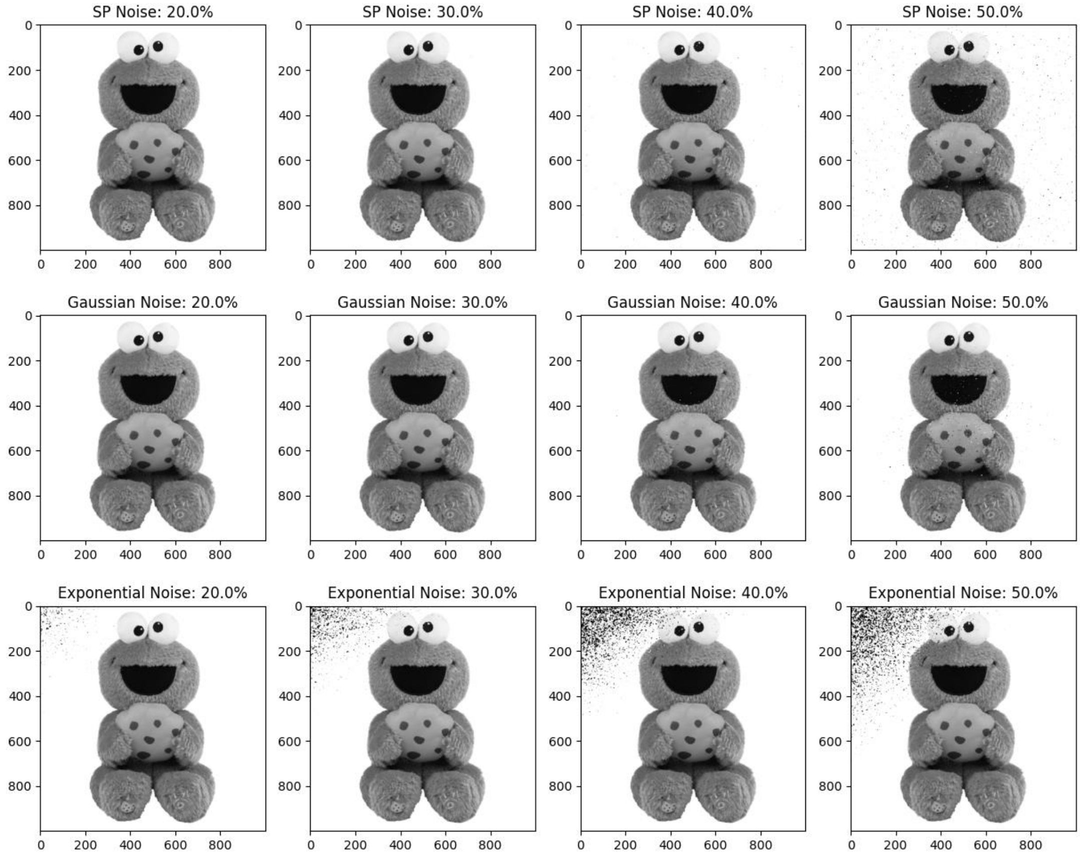


Figure 7-6. Restoring the image by adding a median filter

Problem 8. Conditional Independence

Part a. To prove independence, we need to check if the probability of both events occurring simultaneously is equal to the product of their individual probabilities.

$\boldsymbol{A}$: The event that the first die shows a 6.

$\boldsymbol{B}$: The event that the second die shows a 5.

The probability of A occurring (the first die shows a 6) is:

$$P(A) = \frac{1}{6}$$

The probability of B occurring (the second die shows a 5) is:

$$P(B) = \frac{1}{6}$$

The probability of both events A and B occurring (the first die shows a 6 and the second die shows a 5) is (there are $6 \times 6 = 36$ possible outcomes):

$$P(A \cap B) = \frac{1}{36}$$

Now, check if $P(A \cap B) = P(A)P(B)$:

$$P(A)P(B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36} = P(A \cap B)$$

Part b. Now we want to examine whether these events remain independent given that the sum of the two dice is greater than 10.

$\boldsymbol{C}$: The event that the sum of the two dice is greater than 10.

The possible outcomes when two dice are rolled that give a sum greater than 10 are:

$$(5,6), \qquad (6,5), \qquad (6,6)$$

This gives us 3 outcomes out of a total of 36 possible outcomes, so:

$$P(C) = \frac{3}{36} = \frac{1}{12}$$

$\boldsymbol{P(A \mid C)}$: Probability that the first die shows a 6 given that the sum is greater than 10

Out of these 3 outcomes, 2 have the first die showing a 6, so:

$$P(A \mid C) = \frac{2}{3}$$

$\boldsymbol{P(B \mid C)}$: Probability that the second die shows a 5 given that the sum is greater than 10

Similarly, out of the 3 outcomes in C, only 1 has the second die showing a 5, so:

$$P(B \mid C) = \frac{1}{3}$$

The event $A \cap B$ (both the first die shows a 6 and the second die shows a 5) occurs only in the outcome $(6,5)$. Since $(6,5)$ is one of the 3 outcomes in C, we have:

$$P(A \cap B \mid C) = \frac{1}{3}$$

Now, check if $P(A \cap B \mid C) = P(A \mid C)P(B \mid C)$:

$$P(A \mid C)P(B \mid C) = \frac{2}{3} \times \frac{1}{3} = \frac{2}{9} \neq P(A \cap B \mid C)$$

So, the events A (first die shows a 6) and B (second die shows a 5) are **not independent** given that the sum of the two dice is greater than 10.

Part c. Let's denote the three events involving a single die as E_1, E_2, and E_3, where:

$$E_1 \subseteq E_2 \subseteq E_3$$

This means that if E_1 occurs, then E_2 and E_3 also occur; if E_2 occurs, then E_3 also occurs.

In probability theory, if $E_1 \subseteq E_2 \subseteq E_3$, conditioning on the largest event, E_3, can often render E_1 and E_2 independent. The intuition is that conditioning on E_3 (which is the most inclusive event) provides a complete context within which the occurrence of E_1 does not affect the occurrence of E_2, and vice versa, because both are contained within E_3.

- **Condition on E_3**: $E_1 \subseteq E_2 \subseteq E_3$ implies that $E_1 \cap E_2 = E_1$ (since E1E_1E1 is the smaller set). Given E_3, we need to check if:
  $$P(E_1 \cap E_2 \mid E_3) = P(E_1 \mid E_3) . P(E_2 \mid E_3)$$
  Since E_3 encompasses both E_1 and E_2, conditioning on E_3 often neutralizes dependencies within E_3, making the smaller subsets potentially independent when conditioned on E_3.

- **Condition on E_2**: We check if E_1 and E_3 become independent when conditioned on E_2:
  $$P(E_1 \cap E_3 \mid E_2) = P(E_1 \mid E_2) . P(E_3 \mid E_2)$$
  E_2 contains E_1 but is also part of E_3. Conditioning on E_2 could make E_1 and E_3 appear more independent if E_2 sufficiently captures the overlap between them.

- **Condition on E_1**: We check if E_2 and E_3 become independent when conditioned on E_1:
  $$P(E_2 \cap E_3 \mid E_1) = P(E_2 \mid E_1) . P(E_3 \mid E_1)$$
  Since E_1 is the smallest event, conditioning on E_1 might not sufficiently capture the behavior of E_2 and E_3 unless they are structured such that their independence is preserved when conditioned on E_1.

Part d. Aside from the cases already discussed, another fundamental scenario for examining conditional independence is the **collider structure**, also known as V-structure or explaining away.

In a collider structure, two events (or variables), say A and B, are conditionally dependent **when conditioned on a third event** C, where CCC is influenced by both A and B. The structure is typically represented as:

$$A \rightarrow C \leftarrow B$$

In this setup:

- A and B are **independent** in the absence of information about C.
- Conditioning on C introduces a **dependence** between A and B.

Example: Consider an example where:

- $\boldsymbol{A}$**:** It is raining.
- $\boldsymbol{B}$**:** The sprinkler is on.
- $\boldsymbol{C}$**:** The grass is wet.

If we know that the grass is wet (C), discovering that it is raining (A) makes it less likely that the sprinkler (B) was also on, since rain alone could explain the wet grass.

Problem 9. Limit Distributions

The Poisson distribution is defined by the probability mass function ($\lambda > 0$):

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \qquad k = 0, 1, 2, \dots$$

To be a valid probability function, the Poisson distribution must satisfy two properties:

1. **Non-negativity:** For any $k \geq 0$, $\lambda > 0$, we have $e^{-\lambda} > 0$, and we see that

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} > 0 \rightarrow P(X = k) > 0$$

2. **Normalization:** We need to show that the sum of $P(X = k)$ over all k equals 1.

$$\sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$

The series $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$ is the Taylor series expansion for e^{λ}, so:

$$\sum_{k=0}^{\infty} P(X = k) = e^{-\lambda} e^{\lambda} = 1$$

So, we have proven that the Poisson distribution satisfies the conditions for a **valid probability distribution**.

Now, let's move on to the Binomial distribution. The Binomial distribution describes the probability of k successes in n independent trials, each with success probability p. The Binomial probability mass function is given by:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \qquad k = 0, 1, 2, \dots, n$$

As $n \to \infty$ and $p \to 0$ such that $np = \lambda$, we approximate each term in the formula.

- When n is large and k is fixed, we can approximate $\binom{n}{k}$ as

$$\binom{n}{k} \approx \frac{n^k}{k!}$$

- Since $np = \lambda$, we have

$$p^k = \left(\frac{\lambda}{n}\right)^k = \frac{\lambda^k}{n^k}$$

- Using the limit $\lim\limits_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$, we can approximate $(1 - p)^{n-k}$ as

$$(1 - p)^{n-k} = \left(1 - \frac{\lambda}{n}\right)^{n-k} \approx e^{-\lambda}$$

Substituting these approximations into the Binomial probability mass function

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} = \frac{n^k}{k!} \frac{\lambda^k}{n^k} e^{-\lambda} \approx \binom{n}{k} p^k (1 - p)^{n-k}$$

Thus, as $\boldsymbol{n} \to \infty$ and $\boldsymbol{p} \to \boldsymbol{0}$ such that $\boldsymbol{np} = \boldsymbol{\lambda}$, the Binomial distribution converges to the Poisson distribution:

$$\boldsymbol{P(X = k) = \binom{n}{k} p^{k}(1 - p)^{n-k} \to X \sim Binomial(p; n, k)}$$

Problem 10. Memorylessness

A geometric random variable X with success probability p represents the number of trials needed to get the first success. Its probability mass function (PMF) is:

$$P(X = k) = p\,(1-p)^k\,, \qquad k = 1,2,3,\ldots$$

The probability that $X > n$, the probability that the first success occurs after the n-th trial, is

$$P(X > k) = (1-p)^k$$

Now we want to find the conditional probability $P(X > n + k - 1 \mid X > n - 1)$.

$$P(X > n + k - 1 \mid X > n - 1) = \frac{P(X > n + k - 1\,, X > n - 1)}{P(X > n - 1)} = \frac{P(X > n + k - 1)}{P(X > n - 1)} = \frac{(1-p)^{n+k-1}}{(1-p)^{n-1}}$$

$$= (1-p)^k = P(X > k)$$

$$\rightarrow \boldsymbol{P(X > n + k - 1 \mid X > n - 1) = P(X > k)}$$

Problem 11. Transformation of Random Variables

Let $F_Y(y)$ be the cumulative distribution function (CDF) of Y. By definition,

$$F_Y(y) = P(Y \leq y) = P(g(X) < y)$$

If $g(.)$ is increasing, then $g(X) \leq y$ is equivalent to $X \leq g^{-1}(y)$. Therefore

$$F_Y(y) = P\left(X \leq g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right)$$

To find the PDF $f_Y(y)$, we differentiate $F_Y(y)$ with respect to y:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(g^{-1}(y)\right) = f_X\left(g^{-1}(y)\right)\frac{dg^{-1}(y)}{dy}$$

Now we want to find the expectation of Y in terms of x and y.

$$E_Y(Y) = \int_{-\infty}^{+\infty} y f_Y(y) dy = \int_{-\infty}^{+\infty} y f_X\left(g^{-1}(y)\right)\frac{dg^{-1}(y)}{dy} dy$$

Change of Variables: $\begin{cases} y = g(x) \ \text{ or } \ x = g^{-1}(y) \\ dy = g'(x)dx \text{ or } dx = \left(g^{-1}(y)\right)' dy \end{cases}$

$$\rightarrow E_Y(Y) = \int_{-\infty}^{+\infty} g(x) f_X(x) dx = E_X\left(g(X)\right) = E_X(Y)$$

So, we can say:

$$\boldsymbol{E_Y(Y) = E_X\left(g(X)\right) = E_X(Y)}$$

Problem 12. What are AutoEncoders doing here?

We are given that the latent space Z is drawn from an exponential distribution with rate parameter $\lambda = En(X)$, i.e.,

$$Z \mid X \sim Exp(En(X))$$

The exponential distribution has a probability density function (PDF) given by:

$$f_Z(z \mid X) = En(X).exp(-En(X).z)\,, \qquad z \geq 0$$

The exponential distribution has a well-known expected value (mean), which is given by the inverse of the rate parameter λ. Specifically, for an exponential distribution with rate parameter λ, the expected value is:

$$E(\,Z \mid X\,) = \frac{1}{\lambda} = \frac{1}{En(X)} \to E(\,Z \mid X\,) = \frac{1}{En(X)}$$

Based on the **iterated expectation theorem**, we can write:

$$E(Z) = E_X\big(E(Z|X)\big) = E_X\left(\frac{1}{En(X)}\right)$$

The distribution of X is standard normal $\mathcal{N}(0,1)$. Therefore, we will have:

$$X \sim \mathcal{N}(0,1) \to \; f_X(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

$$E(Z) = E_X\left(\frac{1}{En(X)}\right) = \int_{-\infty}^{+\infty} \frac{1}{En(X)} f_X(x)dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \frac{\exp\left(-\frac{x^2}{2}\right)}{En(X)}dx$$

$$\boldsymbol{E(Z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \frac{\exp\left(-\frac{x^2}{2}\right)}{En(X)}dx}$$

Problem 13. Plants vs. Enemies

The farm is represented as the area under the line $X + Y = 1$, which forms a right triangle with vertices at $(0,0)$, $(1,0)$, and $(0,1)$.

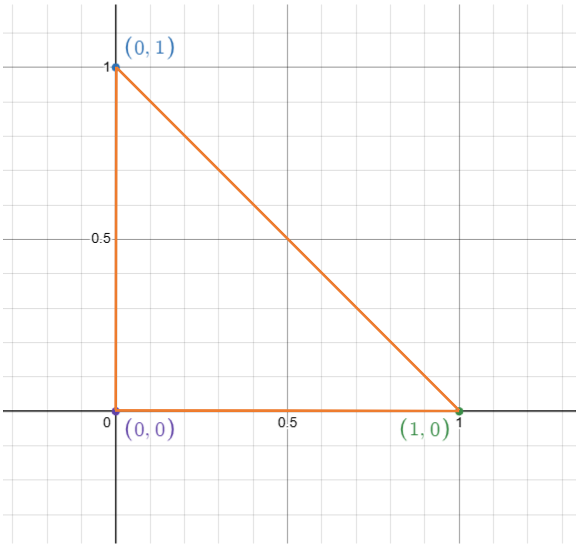


Figure 13-1. The farm area

To minimize the probability of encountering the enemy, the farmer should:

1. Divide the farm area into two regions.
2. Choose the region where the probability of the enemy's presence is lowest.

If we choose the line $X = x_0$, the two regions are divided as follows:

- $Region\ 1: X < x_0\ , 0 < Y < 1 - X \rightarrow Area_{Region_1} = \frac{x_0(2-x_0)}{2}$
- $Region\ 2: X \geq x_0\ , 0 < Y < 1 - X \rightarrow Area_{Region_2} = \frac{(1-x_0)^2}{2}$

Given that the enemy's X-coordinate follows a uniform distribution from 0 to 1, we can say that in region 1, there is a probability of x_0 of encountering the enemy, and in region 2, there is a probability of $1 - x_0$.

- $P\big((x,y) \in Region_1\big) = x_0$
- $P\big((x,y) \in Region_2\big) = 1 - x_0$

If the line $X = x_0 = \frac{1}{2}$ is chosen, the enemy will fall into regions 1 and 2 with equal probability; however, the area of region 1 will be three times the area of region 2.

- $Area_{Region_1} = \frac{\frac{1}{2}\left(2-\frac{1}{2}\right)}{2} = \frac{3}{8} \rightarrow P\big((x,y) \in Region_1\big) = \frac{1}{2}$
- $Area_{Region_2} = \frac{\left(1-\frac{1}{2}\right)^2}{2} = \frac{1}{8} \rightarrow P\big((x,y) \in Region_2\big) = \frac{1}{2}$

As a result, **region 1** is more suitable because, relative to its larger area, fewer enemies will affect it.

$$X < \frac{1}{2}\ , \qquad 0 < Y < 1 - X$$

Problem 14. Random Genes (Optional) (Simulation_Problems.ipynb file)

To implement this part, we first generate a 10 by 100 random uniform matrix. This matrix represents that in 100 experiments, we define 10 random uniform numbers (one number for each gene). Then, in each experiment, we sort the 10 numbers corresponding to the genes and assign them to the genes 1 to 10 in ascending order.

Consider gene number 5. For this gene, we have 100 random uniform numbers. The distribution of these 100 numbers can represent the frequency distribution of gene 5. As shown in figure 14-1, this gene has a normal distribution of frequency change.

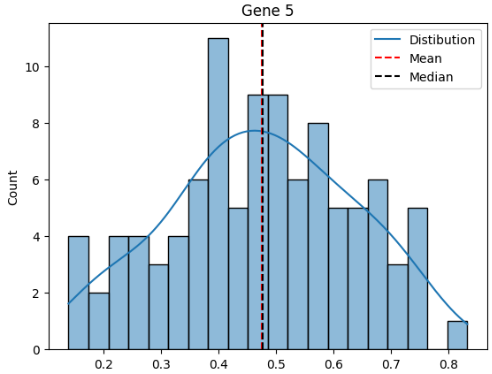


Figure 14-1. The frequency distribution of gene 5

Figure 14-2 shows the frequency distribution of change for each of the genes.

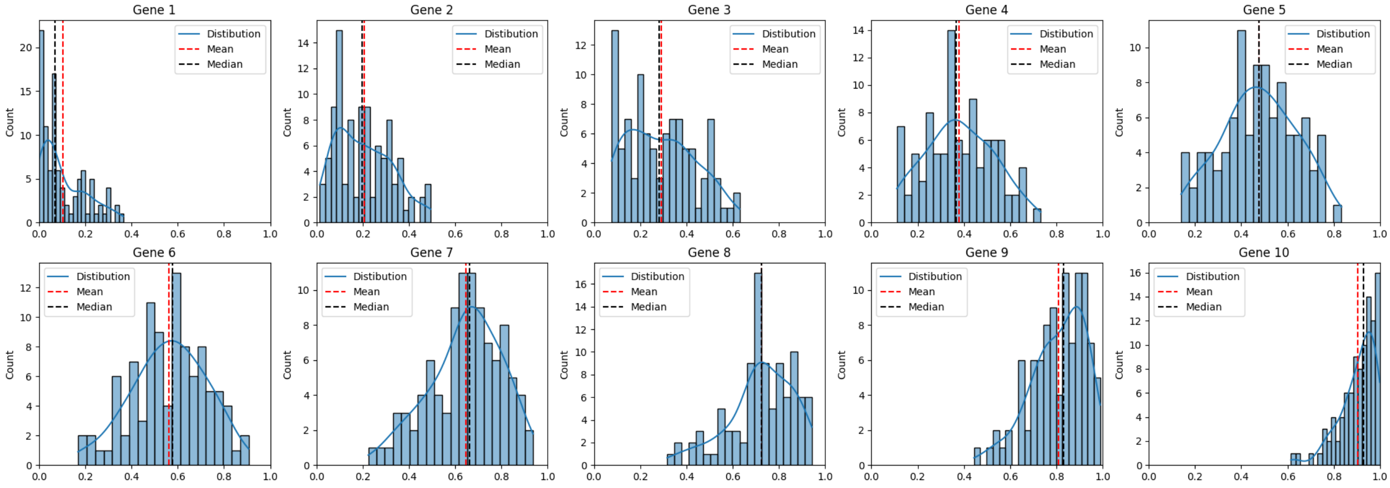


Figure 14-2. The frequency distribution of gene 5

As can be seen, gene 1 has a lower mean frequency of change compared to genes with higher numbers. In other words, the mean frequency of change increases as the gene number increases.