School of Information Technology and Engineering

Master of Science in Artificial Intelligence

Probabilistic Graphical Models

"H1 probabilistic distributions"

Name:- Ruth Tamiru

ID:- GSR 7345/16

Submitted on March 2024

Probaility Distrubutions Lab Report

1. Objective

The objective of the question is to assess our understanding of conditional probability in the context of dice rolls. It tests our ability to calculate probabilities when additional information (conditions) is known about the events.

Tools Required

The only tool we need to know to solve this problem was the formula of the conditional probability which is given by:

P(B|A) = P(A and B) / P(A)

Where:

- P(B|A) is the conditional probability of B given A.
- P(A and B) is the joint probability of both A and B happening.
- P(A) is the probability of event A occurring.

Tasks

Implementing the above concept to each of the questions resulted in:

P(A=3, B=3 | S=6) = 0.027777777777777776
P(S=6 | A=3, B=3) = 1.0
P(A=1, B=1 | S=6) = 0.0
P(A=1|S=2) = 0.16666666666666666
P(A=3|S=2) = 0.0
P(A=2|S=6) = 0.16666666666666666
P (S=12) = 0.027777777777777776
P (S=6) = 0.1388888888888889

Visualization and insights

The very good insight i have found is its real world application. I was just assuming a simple example of our grades and given that we weren’t able to attend mid exam and some assignments our probability to score a grade of “A” will approach to 0..

2. Objective

The objective of this question is to introduce the concept of expected value with a probabilistic approach in the context of random events (dice rolls). It specifically asks to calculate the average number of rolls required for a fair N-sided die to land on each of its faces at least once.

Tools Required

This problem is a famous problem having the name of coupon collector's problem. It deals with the expected number of trials needed to collect a complete set of unique items.
Here we have a collection of N distinct coupons(In our case die). Then we randomly draw coupons one by one with replacement(In our case we throw the die and record the number) finally we need the expected number of trials required to collect all N coupons at least once(In our case to find the faces if the dies at least once).
The tool we need is only the formula which will be discussed below in the case of coupon collector

Tasks

1. **Consider the First Coupon:** The first draw can land on any of the N coupons with equal probability (1/N).
2. **Probability of Missing a Specific Coupon:** After the first draw, there's a (N-1)/N probability that you haven't collected a specific coupon.
3. **Expected Number of Draws for the First Unique Coupon:** Since each draw has an equal chance (1/N) of being the first time you collect a specific missing coupon, the expected number of draws for the first unique coupon is approximately N. (This is a simplification, as the probability of missing a coupon decreases slightly with each draw.)
4. **Conditional Probability for Subsequent Coupons:** Once you have the first unique coupon, you need to repeat the process for the remaining (N-1) coupons. However, the probability of missing a specific coupon now depends on the number of coupons you already have.
5. **Recursive Approach:** You can use a recursive approach to calculate the expected number of additional draws for each remaining coupon, considering the conditional probability of missing it based on the number already collected.
6. **Summation:** The total expected number of draws (E) is the sum of the expected number of draws for each coupon, including the first unique one (approximately N) and the expected additional draws for the remaining coupons (calculated recursively).
   Formula (Approximation):

   The expected number of draws (E) can be approximated by:

   $E \approx N * \ln(N) + 0.57722$ (Euler's constant)
   In my experiment the result was:
      Estimated expected number of rolls for 7 sides: 18.231
      Theoretical expected number of rolls: 24.5

Visualization and insights

The very good insight i have found is the expected number of rolls increases logarithmically with the number of dies(N). This means that while throwing a small number of dies might take only a few triasl, collecting a much larger set will require a proportionally smaller increase in the average number of trials.

3. Objective

The objective of this question is to analyze the dependency between binary variables a, b, and c based on a given joint probability distribution and demonstrate this using the marignal dependence and conditional independence concept.

Tools Required

We need to know about the concepts of Marginal Dependence and Conditional Independence which is stated below:

Marginal Dependence: That a and b are not statistically independent, meaning their marginal probabilities (considering each alone) don't fully determine the joint probability of them occurring together (considering both simultaneously).

Conditional Independence: That a and b become independent when conditioned on c, meaning knowing the value of c eliminates any dependence between a and b.

Tasks

1. Calculate marignal probabilities
2. Check for marginal independence
3. Calculate conditional joint probability
4. Check for conditional independence

My result becomes:

a and b are marginally dependent (p(a,b) != p(a)p(b) for c=0)
a and b are marginally dependent (p(a,b) != p(a)p(b) for c=1)
a and b are conditionally independent given c=0 (p(a,b|c=0) = p(a|c=0)p(b|c=0))
a and b are conditionally independent given c=1 (p(a,b|c=1) = p(a|c=1)p(b|c=1))

Visualization and insights

The very good insight i have found is c might represent an unobserved or hidden variable that influences both a and b. Knowing the value of c effectively removes the influence of this hidden variable, making a and b appear independent from each other.

4. Objective

The objective of this question is to explore the relationship between two propositions for conditional independence of random variables X and Y given a third variable Z and to get used to manipulate expressions using basic probability definations.

Tools Required

Knowdledge of basic probability definations is the basic tool we need.

Tasks

A implies B if $P(X|Y,Z) = P(X|Z)$ for all values of X,Y and Z (where $P(Y,Z) \neq 0$), then we can show that $P(X,Y|Z) = P(X|Z)P(Y|Z)$.

Using the definition of conditional probability: $P(X,Y|Z) = P(X \cap Y \mid Z) / P(Z)$

Since $P(Y,Z) \neq 0$ (given in statement A), we can rewrite: $P(X,Y|Z) = P(X \cap Y \mid Z) / P(Z) = P(X \mid Y, Z) * P(Y \mid Z)$

From the given condition in statement A: $P(X,Y|Z) = P(X \mid Z) * P(Y \mid Z)$

Therefore, if (A) holds true, then (B) also holds true.
2) B implies A

$P(X,Y|Z) = P(X|Z)P(Y|Z)$

Break down the conditional probability on the left side using the definition: $P(X \cap Y \mid Z) / P(Z) = P(X|Z)P(Y|Z)$

Since we're assuming (B) holds true, we can substitute: $P(X \cap Y \mid Z) / P(Z) = P(X|Z)P(Y|Z)$

Now, we want to isolate $P(X|Y,Z)$ on the left side. To achieve this, multiply both sides by $P(Z)$: $P(X \cap Y \mid Z) = P(X|Z)P(Y|Z) * P(Z)$

Since multiplication of probabilities is commutative (order doesn't affect the product), we can rearrange the terms on the right side: $P(X \cap Y \mid Z) = P(Z) * P(X|Z)P(Y|Z)$

Result

f $P(Z) = 0$, the entire right side becomes zero. This aligns with statement (A), where $P(Y,Z) = 0$ (since Z represents the event where Y and Z occur together). In this case, the probability of both X and Y happening given Z is zero, regardless of the relationship between X and Y.

If $P(Z) \neq 0$ (Z has a non-zero probability), then to satisfy the equation: $P(X \cap Y \mid Z) = P(Z) * P(X|Z)P(Y|Z)$

The only way this holds true is if:

P(X|Y,Z) = P(X ∩ Y | Z) / P(Z) (i) Equation (i) represents the definition of conditional probability of X given Y and Z. So, if (B) holds true (meaning the product on the right side exists and is non-zero), then (A) must also hold true for the equation to be balanced.

5. Objective

The objective of this question is to code the formulas for univariate and bivariate gaussian disributions from scratch and to visualize the distribution for better understanding .

Tools Required

We really need to know the formula and the algorithm behind univariate and bivariate gaussian distributions which is stated below:

For Univariate

f(x) = (1 / √(2πσ^2)) * exp(-(x - μ)^2 / (2σ^2))

where:

- x: The value of the random variable
- μ: Mean of the distribution
- σ^2: Variance of the distribution
- √(2πσ^2): Normalization constant to ensure the total area under the curve integrates to 1 (π is the mathematical constant pi)
- exp( ): Exponential function

For Bivariate

f(x, y) = (1 / (2π√(det(Σ)))) * exp(-0.5 * [(x - μ_x), (y - μ_y)] * Σ^-1 * [(x - μ_x), (y - μ_y)]^T)

where:

- x, y: Values of the random variables X and Y
- μ: Mean vector, denoted as [μ_x, μ_y] (mean of X and mean of Y)
- Σ: Covariance matrix, a 2x2 matrix representing the covariance between X and Y
- det(Σ): Determinant of the covariance matrix
- Σ^-1: Inverse of the covariance matrix
- [(x - μ_x), (y - μ_y)]^T: Transpose of the vector [x - μ_x, y - μ_y]

Tasks

Write the above formulas using the appropriate python technique by defining local functions and compare the value with the one done with imported libraries.

Finally the result becomes:
Univariate Gaussian PDF at 7 : 0.12098536225957168
Bivariate Gaussian PDF at 4 , 2.5 : 1.555168909296921

Visualization and insights

The very good insight i have found is that the role of numpy library is essential for the locally defined functions because wile i was working without numpy the result for bivariate was 0.173293656793068 even though the correct answer was 1.555168708929694

6. Objective

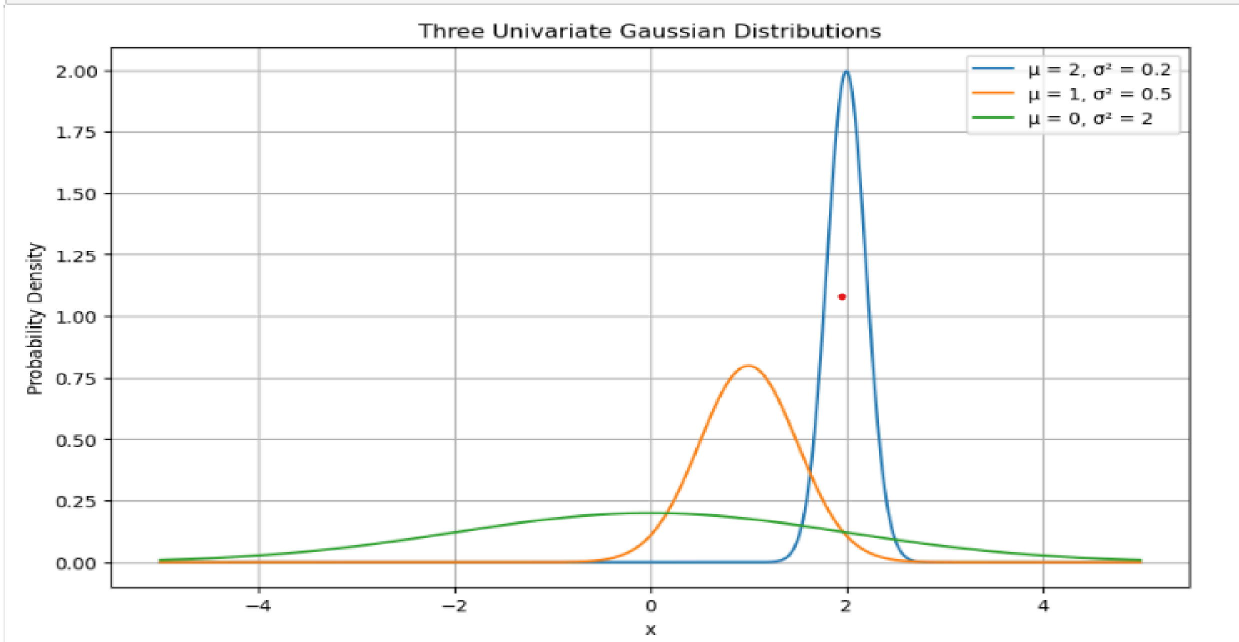
The objective of this question is to investigate the marginal distribution of a specific variable (X) in a bivariate Gaussian distribution using numerical integration and finally to plot a 3D for a bivariate gaussian with the specified covariance matrix.

Tasks

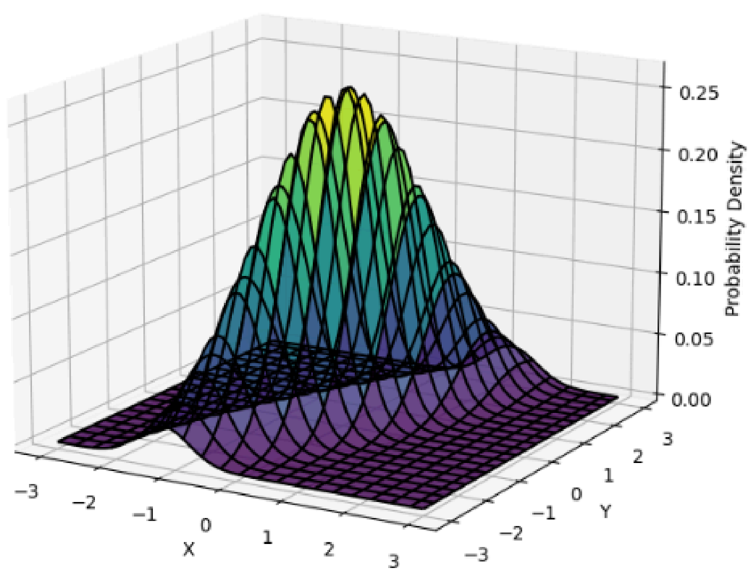
1. **Approximate p(x):** Calculate the marginal probability density function (PDF) of X by integrating the joint PDF of the bivariate Gaussian distribution over all possible values of Y. Numerical integration techniques will be used for this approximation.
2. **Plot the Result:** Visualize the obtained p(x) to understand the distribution of X.
3. **Shape Analysis:** Compare the shape of p(x) with my expectations

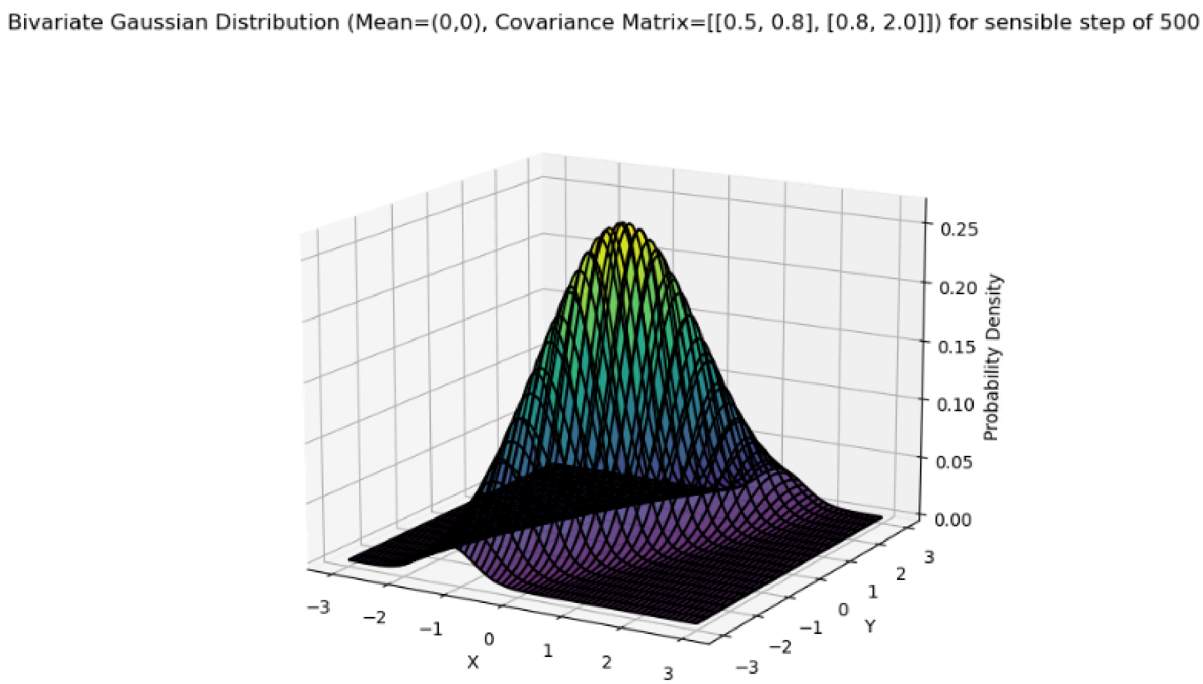
Visualization and insights

Since the original bivariate Gaussian distribution has a mean vector of (0, 0) and a positive covariance between X and Y, we expect the marginal distribution of X (p(x)) to also be a Gaussian distribution centered around 0.
The shape we have found is indeed a Gaussian distribution centered around 0, but slightly wider compared to the original bivariate distribution due to the positive covariance between X and Y.



Bivariate Gaussian Distribution (Mean=(0,0), Covariance Matrix=[[0.5, 0.8], [0.8, 2.0]])





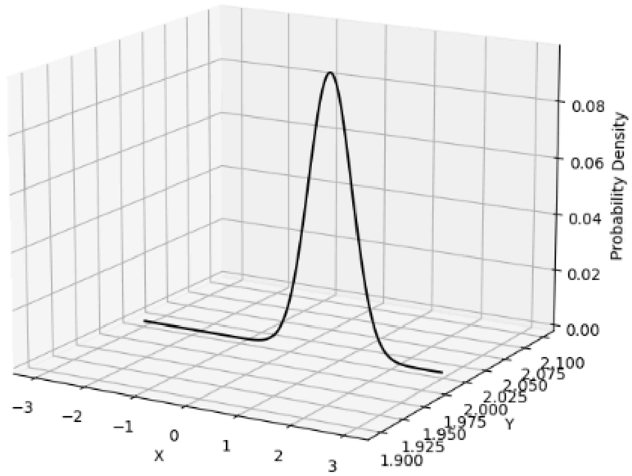
7. Objective

The same objective as that of question number 6.

Visualization and insights

Since the original bivariate Gaussian distribution has a positive covariance between X and Y, we expect the conditional distribution p(x|y=2.0) to be another Gaussian distribution.
The shape is indeed a Gaussian distribution centered around a shifted mean compared to the original distribution (due to the influence of Y). The area under the curve is expected to be 1, signifying the sum of probabilities for all possible values of X given Y=2.0.

Bivariate Gaussian Distribution (Mean=(0,0), Covariance Matrix=[[0.5, 0.8], [0.8, 2.0]]) for sensible step of 500 and p(x|y=2.0)

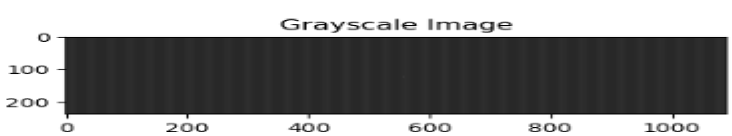


8. Objective

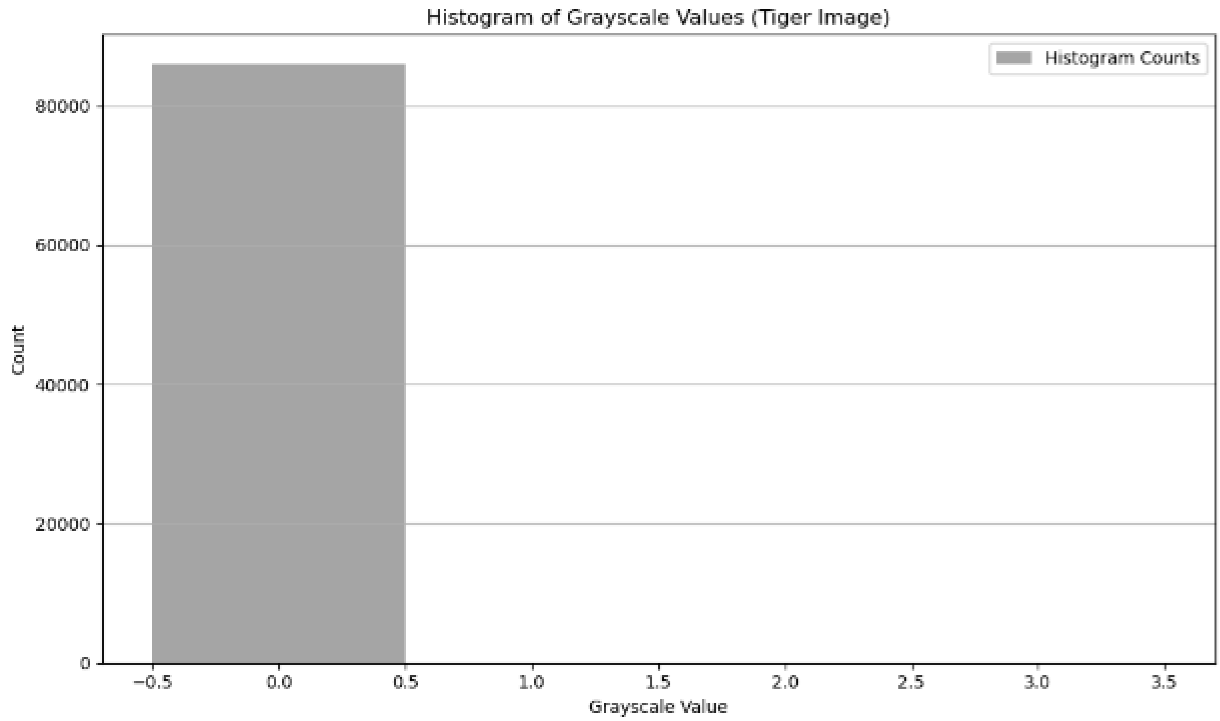
The objective of this question is to display the image from a text file, analyze pixel intensities, generate random images,and estimate samples for exact match.

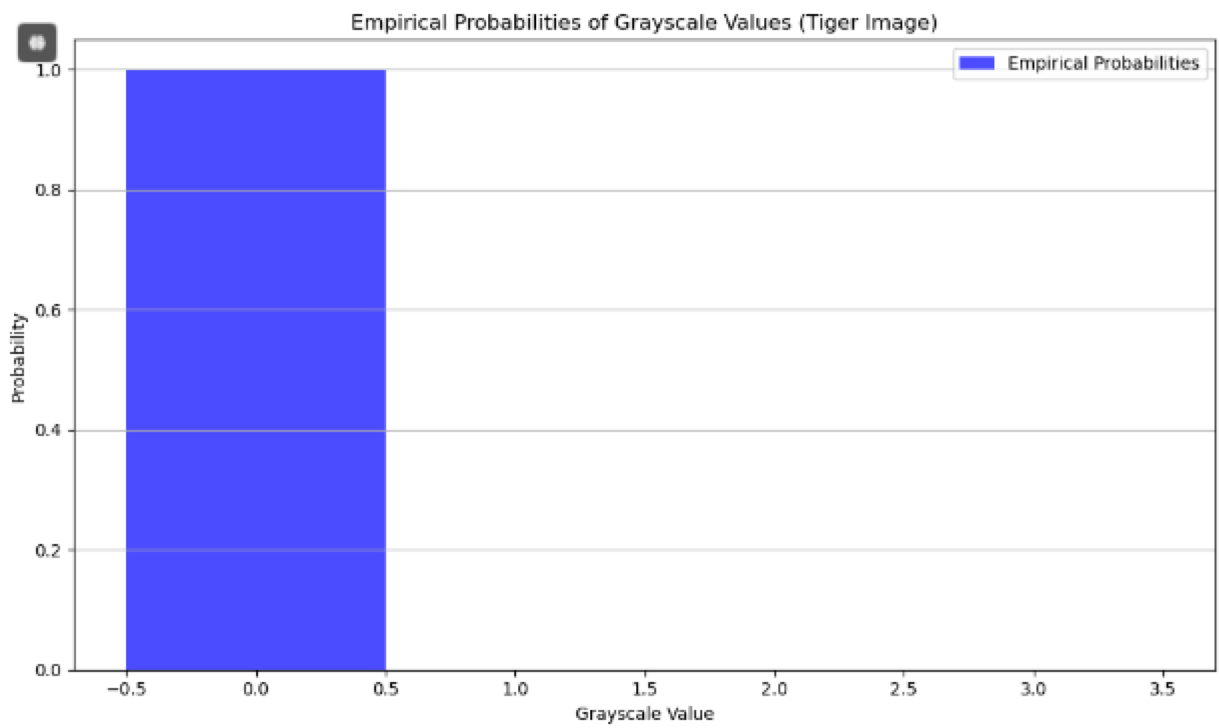
Visualization and insights

Here is the image format for the text given named as tiger.txt.:

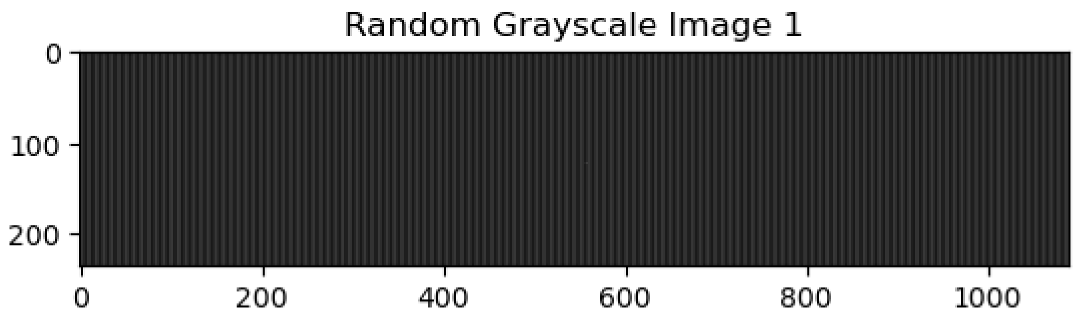


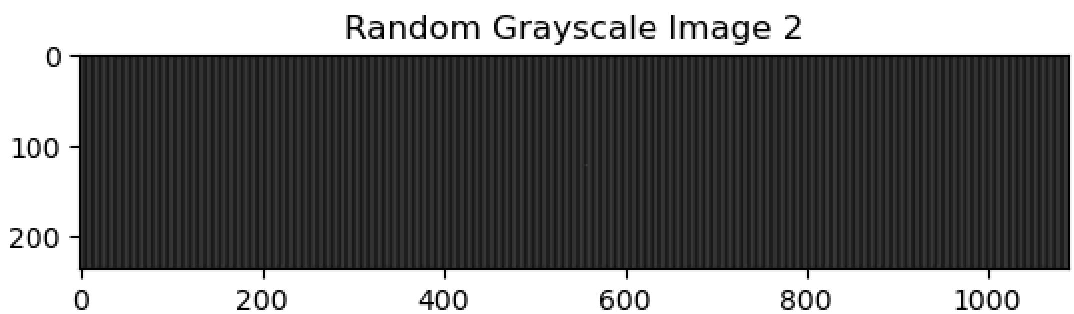
Bar plots for the histogram counts for the grey values of the tiger image

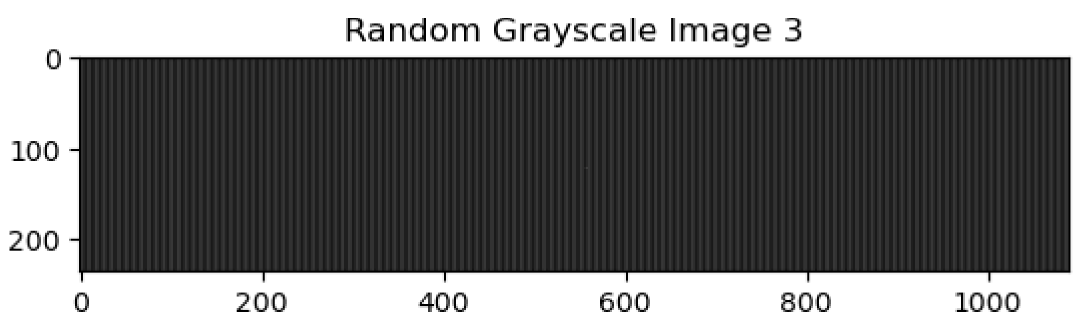




Random generated images are far from the tiger image as shown below







But we can estimate the average number of samples needed to get the exact tiger image using the concept of harmonic series. Approach:

1. Total Number of Images: (256 ^ (236*364)) gives us a large number
2. Probability of getting the tiger image: 1/ total images
3. Harmonic series can be used when the probability of success is low which states that Average trials ≈ 1/ probability(tiger) ≈ total images

calculating the exact total number of images is computationally infeasible. However we can estimate the magnitude. Assuming 256 grayscale values per pixel Total images ≈ (256 ^ (236 *364) ≈ 256^85896