College of Engineering Trivandrum

Department of Electronics and Communication Engineering

ECL332 Communication Engineering Laboratory

Experiment No. 1
Probability and Random Variables – Gaussian Distribution

Name: _____   Date: _____

Roll No.: _____   Batch: _____

Learning Objectives

After completing this experiment, the student will be able to:

- Generate Gaussian (normal) random variables in Python using `NumPy`.
- Visualise and interpret probability density functions (PDFs) and cumulative distribution functions (CDFs) using histograms and plots.
- Verify that the sum of independent Gaussian random variables is also Gaussian and relate the change in variance to the spread of the histogram.
- Estimate probabilities empirically from simulated data by counting samples in specified intervals.
- Experimentally verify the *68–95–99.7 rule* for a Gaussian random variable and compare empirical probabilities with theoretical values.
- Relate the effect of sample size on histogram smoothness and probability estimates, and interpret this behaviour in terms of the law of large numbers.

Theory

Gaussian (Normal) Random Variable

A Gaussian (normal) random variable X with mean μ and variance σ^2 has the probability density function (PDF)

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{1}$$

The graph of this PDF is a bell-shaped curve centred at μ, with spread determined by σ:

- μ shifts the centre of the bell (location).

- σ controls the width/spread of the bell (larger σ gives a wider, flatter curve).

The cumulative distribution function (CDF) of X is defined as

$$F_X(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f_X(t)\,dt. \tag{2}$$

The CDF is a non-decreasing function that approaches 0 as $x \to -\infty$ and 1 as $x \to +\infty$. The slope of the CDF at a point is related to the value of the PDF at that point (steepest where the PDF is largest).

Sum of Independent Gaussian Random Variables

If X and Y are independent Gaussian random variables with

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2), \qquad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2),$$

then their sum

$$Z = X + Y$$

is also Gaussian, with

$$Z \sim \mathcal{N}(\mu_X + \mu_Y,\ \sigma_X^2 + \sigma_Y^2).$$

In this experiment, X and Y have the same distribution $\mathcal{N}(r, 1)$, where r is the roll number. Therefore

$$Z = X + Y \sim \mathcal{N}(2r, 2).$$

The variance of Z is larger than that of X or Y, so its PDF (histogram) is wider.

Empirical PDF, CDF and Histograms

In practice, we do not know the PDF exactly; instead, we observe a finite number of samples. Given N samples of X:

$$x_1, x_2, \ldots, x_N,$$

we can:

- Construct a **histogram** by dividing the x-axis into bins and counting how many samples fall into each bin.

- If we normalise the histogram (e.g. using `density=True` in `plt.hist`), the bar heights approximate the underlying PDF.

- An empirical CDF can be obtained by accumulating the (normalised) histogram values or by sorting the samples and plotting the cumulative fraction.

As N increases, the histogram becomes smoother and closer to the true PDF.

68–95–99.7 Rule for the Normal Distribution

For a Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, the following approximate probabilities hold:

$$\mathbb{P}(|X - \mu| \leq \sigma) \approx 0.68,$$
$$\mathbb{P}(|X - \mu| \leq 2\sigma) \approx 0.95,$$
$$\mathbb{P}(|X - \mu| \leq 3\sigma) \approx 0.997.$$

This is known as the *68–95–99.7 rule*. It tells us that:

- About 68% of the probability mass lies within one standard deviation of the mean.

- About 95% lies within two standard deviations.

- Almost all (99.7%) lies within three standard deviations.

In the lab, we verify this rule empirically by generating many samples of X, counting how many samples fall inside the intervals

$$[\mu - \sigma, \mu + \sigma], \quad [\mu - 2\sigma, \mu + 2\sigma], \quad [\mu - 3\sigma, \mu + 3\sigma],$$

and dividing by the total number of samples.

Law of Large Numbers and Empirical Probabilities

Let $X_1, X_2, \ldots, X_N$ be independent samples of a random variable X. If we estimate a probability such as $\mathbb{P}(a \leq X \leq b)$ by

$$\hat{p}_N = \frac{\text{number of samples with } a \leq X_i \leq b}{N},$$

then the **law of large numbers** states that as $N \to \infty$,

$$\hat{p}_N \to \mathbb{P}(a \leq X \leq b)$$

(with high probability).

Pre-Lab Questions

Answer the following numerical problems **before** coming to the lab. Show all important steps.

1) **Basic probability**

   (a) A box contains 5 red balls, 3 blue balls, and 2 green balls. One ball is drawn at random. Find:

   - $\mathbb{P}(\text{red})$

   - $\mathbb{P}(\text{not blue})$

   - $\mathbb{P}(\text{red or green})$

   (b) A fair die is thrown once. Let X denote the outcome. Compute:

   $$\mathbb{P}(X \leq 3), \quad \mathbb{P}(X \text{ is even}), \quad \mathbb{P}(2 \leq X \leq 5).$$

2) **Standard normal (Gaussian) distribution** Let $X \sim \mathcal{N}(0,1)$ (standard normal random variable). Using the standard normal table, compute approximately:

(a) $\mathbb{P}(X \leq 1)$

(b) $\mathbb{P}(-1 \leq X \leq 1)$

(c) $\mathbb{P}(X > 2)$

3) **General normal distribution** Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = 50$ and $\sigma = 5$.

(a) Convert X to a standard normal variable Z and write the relation between X and Z.

(b) Using standard normal tables, estimate $\mathbb{P}(45 \leq X \leq 55)$.

(c) Estimate $\mathbb{P}(X \geq 60)$.

4) **Sum of Gaussian random variables** Let $X \sim \mathcal{N}(10,1)$ and $Y \sim \mathcal{N}(10,1)$ be independent.

(a) Find the distribution (mean and variance) of $Z = X + Y$.

(b) Using this distribution, write an expression for

$$\mathbb{P}(18 \leq Z \leq 22)$$

in terms of the standard normal CDF. (You need not evaluate it numerically, but simplify it up to Z-scores.)

5) **Histogram and empirical PDF** A set of 20 measured values of a random variable X is given below (e.g., amplitudes of a noisy signal):

8, 9, 10, 10, 11, 12, 12, 12, 13, 13, 13, 14, 14, 15, 16, 16, 16, 17, 18, 20.

(a) Choose class intervals of width 2 starting from $[8, 10)$, i.e.,

$$[8, 10), \ [10, 12), \ [12, 14), \ [14, 16), \ [16, 18), \ [18, 20].$$

Construct a frequency table for these intervals.

(b) Compute the *relative frequency* for each class (frequency divided by total number of samples).

(c) Using these classes and relative frequencies, sketch a histogram. Label the

axes clearly.

Experiment 1: Gaussian Random Variables – PDF, CDF and Summation

1) Set the mean r equal to your roll number. Choose the variance $\sigma^2 = 1$.

2) Decide the number of samples N (e.g. $N = 10^5$).

3) Use `np.random.normal` to generate N samples of X and Y with mean r and standard deviation 1.

4) Plot the histograms of X and Y using `plt.hist`. Choose a suitable number of bins.

5) Approximate the PDFs of X and Y from the histograms and plot them.

6) Compute the empirical CDFs of X and Y using the cumulative sum of the PDF values and plot them.

7) Form a new random variable $Z = X + Y$ and plot the histogram (PDF) of Z.

8) Compute the empirical mean and variance of Z using `np.mean` and `np.var`.

9) Compare the empirical mean and variance of Z with the theoretical values $2r$ and 2.

Observations and Calculations

1. Empirical and Theoretical Statistics

Variable	Empirical Mean	Empirical Variance	Theoretical Mean	Theoretical Variance
X			r	1
Y			r	1
$Z = X + Y$			$2r$	2

2. Plots

Draw the following plots:

- Histograms of X and Y.

- PDFs of X and Y.

- CDFs of X and Y.

- Histogram (PDF) of $Z = X + Y$.

Experiment 1: Post Lab Questions

1) In your histogram of X, how did the shape change when you:

  - (i) used a small number of samples (e.g. $N = 1000$), and
  - (ii) used a large number of samples (e.g. $N = 10^5$)?

  Explain why this happens.

2) In `plt.hist`, what is the effect of using `density=True`? How is this related to approximating the PDF from the histogram?

3) Compare the empirical mean and variance of X and Y that you obtained with the theoretical values r and 1.

- Are they exactly the same?
- If not, give two reasons why they differ slightly.

4) For the random variable $Z = X + Y$:

- What are the theoretical mean and variance of Z?
- Is the histogram of Z "wider" or "narrower" than that of X? Explain using the variance.

5) Look at the CDF plots of X and Y.

- How does the slope of the CDF relate to the PDF?
- Around which value of x is the CDF changing most rapidly, and why?

6) If you change the standard deviation from $\sigma = 1$ to $\sigma = 2$ while keeping the same mean:

  - How does the histogram (PDF) of X change?
  - How does the CDF curve change in terms of steepness?

7) You generated X and Y independently. Suppose you plot a scatter plot of (X, Y) pairs.

  - What pattern would you expect to see if X and Y are independent?
  - What pattern might indicate that X and Y are correlated?

8) From your experiment, do you think increasing the number of samples N always improves the match between empirical and theoretical PDFs? Explain, mentioning any computational limitations you observed (e.g. time, memory, plotting).

Experiment 2: Verification of the "68–95–99.7" Rule for Gaussian Distribution

Using the same random variable $X \sim \mathcal{N}(r, 1)$ (where r is your roll number) generated in this experiment:

a) Using a large number of samples (e.g. $N = 10^5$), estimate the following probabilities empirically:

$$\mathbb{P}(|X - r| \leq 1), \qquad \mathbb{P}(|X - r| \leq 2), \qquad \mathbb{P}(|X - r| \leq 3).$$

(Hint: Count the number of samples that satisfy each condition and divide by N.)

b) Theoretically, for a Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, we have approximately:

$$\mathbb{P}(|X - \mu| \leq \sigma) \approx 0.68, \quad \mathbb{P}(|X - \mu| \leq 2\sigma) \approx 0.95, \quad \mathbb{P}(|X - \mu| \leq 3\sigma) \approx 0.997.$$

Compare your empirical probabilities with these theoretical values and tabulate the results.

Experiment 2: Post Lab Questions

3) Describe step-by-step how you estimated $\mathbb{P}(|X - r| \leq 1)$ using your simulated data. (Do not give a formula; explain the counting procedure.)

4) Tabulate the theoretical and empirical probabilities you obtained for:

$$\mathbb{P}(|X - r| \leq 1), \quad \mathbb{P}(|X - r| \leq 2), \quad \mathbb{P}(|X - r| \leq 3).$$

Comment on which one (1-sigma, 2-sigma or 3-sigma) matched the theoretical value most closely in your experiment.

5) Suppose you reduce the number of samples from $N = 10^5$ to $N = 500$.

- What happens to the accuracy of your empirical estimates of these probabilities?

- Relate your observation to the law of large numbers.

6) Out of $N = 10^5$ samples, roughly how many samples would you theoretically expect to lie *outside* the interval $[r - 3, r + 3]$? Compare this with what you actually observed. Is it close? Why or why not?

7) In your own words, explain what the "68–95–99.7 rule" tells you about where most of the probability mass of a Gaussian random variable lies. How is this visible in:

- (i) the PDF plot, and
- (ii) the CDF plot?

8) If instead of $\sigma = 1$ you had $\sigma = 2$, what would be the new intervals corresponding to:
$$\text{(i)} \ \pm 1\sigma, \quad \text{(ii)} \ \pm 2\sigma, \quad \text{(iii)} \ \pm 3\sigma$$
around the mean r? Would the 68–95–99.7 percentages change? Answer and justify.

Concluding Remarks

(*Give briefly what you have understood from the experiment.*)

Student's Signature: _____