Homework 3 - Applied Stochastic Processes

kipngeno koech - bkoech

October 26, 2024

Question 1: Random Vectors and Principal Component Analysis

Reading: Random vectors are fundamental constructs in probability and statistics, allowing researchers and practitioners to analyze relationships among multiple variables simultaneously. Each component of a random vector can represent a different feature or measurement, and the joint distribution encapsulates the uncertainty inherent in those variables.

For instance, consider a random vector X - $X_1, X_2, ..., X_n$ where each X_i is a random variable. The covariance matrix of X plays a crucial role in understanding the linear relationships among the components, guiding decisions in fields such as finance, machine learning, and signal processing. Sampling from random vectors introduces excitement in multivariate analyses, where one can explore properties like independence, marginal distributions, and conditional relationships. Moreover, techniques such as principal component analysis (PCA) leverage the variance structure of these vectors to reduce dimensionality while preserving essential information.

1. **5 points** let $X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$, and $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$ are related by $Y - AX$ where

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

The joint PMF of X is given by:

$$P_X(X) = \begin{cases} (1-p)p^{x_3} & \text{if } X_1 < X_2 < X_3 \\ 0 & \text{otherwise} \end{cases}$$

where $x_1, x_2, x_3 \in \{0, 1, 2, \ldots\}$ and $0 < p < 1$.
Find the joint PMF $P_Y(y)$ of the transformed random vector Y.

$$Y = AX = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 - X_1 \\ X_3 - X_2 \end{pmatrix}$$

$$X_1 = Y_1, X_2 = Y_1 + Y_2, X_3 = Y_1 + Y_2 + Y_3$$

$$P_Y(y) = P_X(A^{-1}y) = P_X\left(\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}\right) = P_X\left(\begin{pmatrix} y_1 \\ y_1 + y_2 \\ y_1 + y_2 + y_3 \end{pmatrix}\right)$$

conditions for $X_1 < X_2 < X_3$ to hold:

$$y_1 < y_1 + y_2 < y_1 + y_2 + y_3$$

$$0 < y_2 < y_3$$

$$P_Y(y) = (1-p)p^{y_1+y_2+y_3} = (1-p)p^{y_1}p^{y_2}p^{y_3}$$

$$P_Y(y) = (1-p)p^{y_1}p^{y_2}p^{y_3}$$

2. You are working as a data analyst for a startup that collects various statistics from users' activities on its platform. The startup wants to reduce the dimensionality of its collected data without losing significant information. Your goal is to apply Principal Component Analysis (PCA) to the dataset to retain as much variance (information) as possible while reducing the dimensionality. This exercise will take you from the conceptual understanding of random vectors and covariance matrices to the practical application of PCA using Python.

Part 1: understanding the covariance Matrix of Random Vectors (12 points)
You are given a random vector $X = [X_1, X_2, X_3, X_4]^T$, representing four features of platforms users. The covariance matrix of this random vector is:

$$\Sigma_x = \begin{bmatrix} 5 & 1.2 & 0.8 & 0.6 \\ 1.2 & 4 & 0.5 & 0.3 \\ 0.8 & 0.5 & 3 & 0.2 \\ 0.6 & 0.3 & 0.2 & 2 \end{bmatrix}$$

Intepretation of the covariance matrix

(a) **(2 points)** What do the diagonal elements of the covariance matrix represent?

The diagonal elements of the covariance matrix represent the variance of the individual features.

(b) **(2 points)** What do the off-diagonal elements signify in terms of the relationship between the features?

The off-diagonal elements signify the covariance between the features.

Random Vector and Variance

(a) **(2 points)** Calculate the total variance of the random vector X.

$$\text{Total Variance} = \text{Trace}(\Sigma_x) = 5 + 4 + 3 + 2 = \mathbf{14}$$

(b) **(2 points)** How would you compute the variance captured by a single feature (e.g, the first feature X_1)?

$$\text{Variance of } X_1 = \Sigma_{11} = \mathbf{5}$$

Eigenvalues and Eigenvectors of the Covariance Matrix

(a) **(2 points)** Calculate the eigenvalues and eigenvectors of the covariance matrix Σ_x by hand
method used to calculate eigenvalues and eigenvectors is eigen decomposition:
The characteristic equation is given by:

$$\det(\Sigma_x - \lambda I) = 0$$

$$\begin{vmatrix} 5-\lambda & 1.2 & 0.8 & 0.6 \\ 1.2 & 4-\lambda & 0.5 & 0.3 \\ 0.8 & 0.5 & 3-\lambda & 0.2 \\ 0.6 & 0.3 & 0.2 & 2-\lambda \end{vmatrix} = 0$$

$$\lambda^4 - 14\lambda^3 + 68.18\lambda^2 - 139.254\lambda + 101.356 = 0$$

$$\lambda_1 = \mathbf{6.20306}, \lambda_2 = \mathbf{3.20619}, \lambda_3 = \mathbf{2.71066}, \lambda_4 = \mathbf{1.88009}$$

To calculate the eigen vectors, we substitute the eigen values into the equation:
The eigen vectors are:
for eigen value $\lambda_1 = 6.20306$

$$\begin{bmatrix} -1.20306 & 1.2 & 0.8 & 0.6 \\ 1.2 & -2.20306 & 0.5 & 0.3 \\ 0.8 & 0.5 & -3.20306 & 0.2 \\ 0.6 & 0.3 & 0.2 & -4.20306 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

The augmented matrix is:

$$\left[\begin{array}{cccccc} -1.20306 & 1.2 & 0.8 & 0.6 & | & 0 \\ 1.2 & -2.20306 & 0.5 & 0.3 & | & 0 \\ 0.8 & 0.5 & -3.20306 & 0.2 & | & 0 \\ 0.6 & 0.3 & 0.2 & -4.20306 & | & 0 \end{array}\right]$$

$$R_1 = \frac{1}{-1.20306} R_1 = \left[\begin{array}{cccccc} 1 & -0.997 & -0.66497 & -0.4987 & | & 0 \\ 1.2 & -2.20306 & 0.5 & 0.3 & | & 0 \\ 0.8 & 0.5 & -3.20306 & 0.2 & | & 0 \\ 0.6 & 0.3 & 0.2 & -4.20306 & | & 0 \end{array}\right]$$

$$R_2 = R_2 - 1.2R_1 = \left[\begin{array}{cccccc} 1 & -0.997 & -0.66497 & -0.4987 & | & 0 \\ 0 & -1.00611 & 1.297965 & 0.89847 & | & 0 \\ 0.8 & 0.5 & -3.20306 & 0.2 & | & 0 \\ 0.6 & 0.3 & 0.2 & -4.20306 & | & 0 \end{array}\right]$$

$$R_3 = R_3 - 0.8R_1 = \left[\begin{array}{cccccc} 1 & -0.997 & -0.66497 & -0.4987 & | & 0 \\ 0 & -1.00611 & 1.297965 & 0.89847 & | & 0 \\ 0 & 1.497965 & -2.57024 & 0.59896 & | & 0 \\ 0.6 & 0.3 & 0.2 & -4.20306 & | & 0 \end{array}\right]$$

$$V_1 = \begin{bmatrix} 0.78907948 \\ 0.16466637 \\ 0.52644338 \\ 0.27036259 \end{bmatrix}$$

for eigen value $\lambda_1 = 1.88009$

$$\begin{bmatrix} 3.11991 & 1.2 & 0.8 & 0.6 \\ 1.2 & 2.11991 & 0.5 & 0.3 \\ 0.8 & 0.5 & 1.11991 & 0.2 \\ 0.6 & 0.3 & 0.2 & 0.11991 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$R_1 = \frac{1}{3.11991} R_1 = \left[\begin{array}{cccccc} 1 & 0.384 & 0.256 & 0.192 & | & 0 \\ 1.2 & 2.11991 & 0.5 & 0.3 & | & 0 \\ 0.8 & 0.5 & 1.11991 & 0.2 & | & 0 \\ 0.6 & 0.3 & 0.2 & 0.11991 & | & 0 \end{array}\right] \quad R_2 = R_2 - 1.2R_1 = \left[\begin{array}{cccccc} 1 & 0.384 & 0.256 & 0.192 & | & 0 \\ 0 & 1.658 & 0.192 & 0.069 & | & 0 \\ 0.8 & 0.5 & 1.11991 & 0.2 & | & 0 \\ 0.6 & 0.3 & 0.2 & 0.11991 & | & 0 \end{array}\right]$$

$$R_3 = R_3 - 0.8R_1 = \left[\begin{array}{cccccc} 1 & 0.384 & 0.256 & 0.192 & | & 0 \\ 0 & 1.658 & 0.192 & 0.069 & | & 0 \\ 0 & 0.192 & 0.915 & 0.046 & | & 0 \\ 0.6 & 0.3 & 0.2 & 0.11991 & | & 0 \end{array}\right] \quad R_4 = R_4 - 0.6R_1 = \left[\begin{array}{cccccc} 1 & 0.384 & 0.256 & 0.192 & | & 0 \\ 0 & 1.658 & 0.192 & 0.069 & | & 0 \\ 0 & 0.192 & 0.915 & 0.046 & | & 0 \\ 0 & 0.069 & 0.046 & 0.005 & | & 0 \end{array}\right]$$

$$R_2 = \frac{1}{1.658}R_2 = \begin{bmatrix} 1 & 0.384 & 0.256 & 0.192 & | & 0 \\ 0 & 1 & 0.116 & 0.042 & | & 0 \\ 0 & 0.192 & 0.915 & 0.046 & | & 0 \\ 0 & 0.069 & 0.046 & 0.005 & | & 0 \end{bmatrix} R_3 = R_3 - 0.192R_2 = \begin{bmatrix} 1 & 0.384 & 0.256 & 0.192 & | & 0 \\ 0 & 1 & 0.116 & 0.042 & | & 0 \\ 0 & 0 & 0.892 & 0.038 & | & 0 \\ 0 & 0.069 & 0.046 & 0.005 & | & 0 \end{bmatrix}$$

$$R_4 = R_4 - 0.069R_2 = \begin{bmatrix} 1 & 0.384 & 0.256 & 0.192 & | & 0 \\ 0 & 1 & 0.116 & 0.042 & | & 0 \\ 0 & 0 & 0.892 & 0.038 & | & 0 \\ 0 & 0 & 0.038 & 0.002 & | & 0 \end{bmatrix} R_3 = \frac{1}{0.892}R_3 = \begin{bmatrix} 1 & 0.384 & 0.256 & 0.192 & | & 0 \\ 0 & 1 & 0.116 & 0.042 & | & 0 \\ 0 & 0 & 1 & 0.043 & | & 0 \\ 0 & 0 & 0.038 & 0.002 & | & 0 \end{bmatrix}$$

$$R_4 = R_4 - 0.038R_3 = \begin{bmatrix} 1 & 0.384 & 0.256 & 0.192 & | & 0 \\ 0 & 1 & 0.116 & 0.042 & | & 0 \\ 0 & 0 & 1 & 0.043 & | & 0 \\ 0 & 0 & 0 & 0.000366 & | & 0 \end{bmatrix} R_2 = R_2 - 0.116R_3 = \begin{bmatrix} 1 & 0.384 & 0.256 & 0.192 & | & 0 \\ 0 & 1 & 0 & 0.037012 & | & 0 \\ 0 & 0 & 1 & 0.043 & | & 0 \\ 0 & 0 & 0 & 0.000366 & | & 0 \end{bmatrix}$$

$$R_1 = R_1 - 0.384R_2 = \begin{bmatrix} 1 & 0 & 0.256 & 0.192 & | & 0 \\ 0 & 1 & 0 & 0.037012 & | & 0 \\ 0 & 0 & 1 & 0.043 & | & 0 \\ 0 & 0 & 0 & 0.000366 & | & 0 \end{bmatrix} R_1 = R_1 - 0.256R_3 = \begin{bmatrix} 1 & 0 & 0 & 0.000366 & | & 0 \\ 0 & 1 & 0 & 0.037012 & | & 0 \\ 0 & 0 & 1 & 0.043 & | & 0 \\ 0 & 0 & 0 & 0.000366 & | & 0 \end{bmatrix}$$

$$v_2 = \begin{bmatrix} 0.51741855 \\ 0.037012 \\ -0.84693231 \\ 0.11692356 \end{bmatrix}$$

for eigen value $\lambda_2 = 2.71066$

$$\begin{bmatrix} 2.28934 & 1.2 & 0.8 & 0.6 \\ 1.2 & 1.28934 & 0.5 & 0.3 \\ 0.8 & 0.5 & 0.289 & 0.2 \\ 0.6 & 0.3 & 0.2 & 0.71066 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$R_1 = \frac{1}{2.28934}R_1 = \begin{bmatrix} 1 & 0.524 & 0.349 & 0.262 & | & 0 \\ 1.2 & 1.28934 & 0.5 & 0.3 & | & 0 \\ 0.8 & 0.5 & 0.289 & 0.2 & | & 0 \\ 0.6 & 0.3 & 0.2 & 0.71066 & | & 0 \end{bmatrix} R_2 = R_2 - 1.2R_1 = \begin{bmatrix} 1 & 0.524 & 0.349 & 0.262 & | & 0 \\ 0 & 0.660 & 0.081 & -0.015 & | & 0 \\ 0.8 & 0.5 & 0.289 & 0.2 & | & 0 \\ 0.6 & 0.3 & 0.2 & 0.71066 & | & 0 \end{bmatrix}$$

$$R_3 = R_3 - 0.8R_1 = \begin{bmatrix} 1 & 0.524 & 0.349 & 0.262 & | & 0 \\ 0 & 0.660 & 0.081 & -0.015 & | & 0 \\ 0 & 0.081 & 0.010 & -0.010 & | & 0 \\ 0.6 & 0.3 & 0.2 & 0.71066 & | & 0 \end{bmatrix} R_4 = R_4 - 0.6R_1 = \begin{bmatrix} 1 & 0.524 & 0.349 & 0.262 & | & 0 \\ 0 & 0.660 & 0.081 & -0.015 & | & 0 \\ 0 & 0.081 & 0.010 & -0.010 & | & 0 \\ 0 & -0.015 & -0.010 & -0.868 & | & 0 \end{bmatrix}$$

$$R_2 = \frac{1}{0.660}R_2 = \begin{bmatrix} 1 & 0.524 & 0.349 & 0.262 & | & 0 \\ 0 & 1 & 0.123 & -0.023 & | & 0 \\ 0 & 0.081 & 0.010 & -0.010 & | & 0 \\ 0 & -0.015 & -0.010 & -0.868 & | & 0 \end{bmatrix} R_3 = R_3 - 0.081R_2 = \begin{bmatrix} 1 & 0.524 & 0.349 & 0.262 & | & 0 \\ 0 & 1 & 0.123 & -0.023 & | & 0 \\ 0 & 0 & 0.001 & -0.001 & | & 0 \\ 0 & -0.015 & -0.010 & -0.868 & | & 0 \end{bmatrix}$$

$$R_4 = R_4 + 0.015R_2 = \begin{bmatrix} 1 & 0.524 & 0.349 & 0.262 & | & 0 \\ 0 & 1 & 0.123 & -0.023 & | & 0 \\ 0 & 0 & 0.001 & -0.001 & | & 0 \\ 0 & 0 & -0.010 & -0.868 & | & 0 \end{bmatrix} R_3 = \frac{1}{0.001}R_3 = \begin{bmatrix} 1 & 0.524 & 0.349 & 0.262 & | & 0 \\ 0 & 1 & 0.123 & -0.023 & | & 0 \\ 0 & 0 & 1 & -1 & | & 0 \\ 0 & 0 & -0.010 & -0.868 & | & 0 \end{bmatrix}$$

$$v_2 = \begin{bmatrix} 0.16328172 \\ -0.98478572 \\ 0.05882726 \\ 0.00869061 \end{bmatrix}$$

for eigen value $\lambda_3 = 3.20619$

$$\begin{bmatrix} 1.79381 & 1.2 & 0.8 & 0.6 \\ 1.2 & 2.79381 & 0.5 & 0.3 \\ 0.8 & 0.5 & 2.20619 & 0.2 \\ 0.6 & 0.3 & 0.2 & 1.20619 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$R_1 = \frac{1}{1.79381}R_1 = \begin{bmatrix} 1 & 0.668 & 0.445 & 0.334 & | & 0 \\ 1.2 & 2.79381 & 0.5 & 0.3 & | & 0 \\ 0.8 & 0.5 & 2.20619 & 0.2 & | & 0 \\ 0.6 & 0.3 & 0.2 & 1.20619 & | & 0 \end{bmatrix} R_2 = R_2 - 1.2R_1 = \begin{bmatrix} 1 & 0.668 & 0.445 & 0.334 & | & 0 \\ 0 & 1.125 & 0.107 & -0.025 & | & 0 \\ 0.8 & 0.5 & 2.20619 & 0.2 & | & 0 \\ 0.6 & 0.3 & 0.2 & 1.20619 & | & 0 \end{bmatrix}$$

$$R_3 = R_3 - 0.8R_1 = \begin{bmatrix} 1 & 0.668 & 0.445 & 0.334 & | & 0 \\ 0 & 1.125 & 0.107 & -0.025 & | & 0 \\ 0 & -0.107 & 0.00038 & -0.00038 & | & 0 \\ 0.6 & 0.3 & 0.2 & 1.20619 & | & 0 \end{bmatrix} R_4 = R_4 - 0.6R_1 = \begin{bmatrix} 1 & 0.668 & 0.445 & 0.334 & | & 0 \\ 0 & 1.125 & 0.107 & -0.025 & | & 0 \\ 0 & -0.107 & 0.00038 & -0.00038 & | & 0 \\ 0 & -0.025 & -0.00038 & 0.00038 & | & 0 \end{bmatrix}$$

$$v_3 = \begin{bmatrix} 0.28804636 \\ 0.04206535 \\ 0.04585165 \\ -0.95559271 \end{bmatrix}$$

(b) **(2 points)** List the eigenvalues in descending order and explain what they represent in terms of variance

$$\lambda_1 = \mathbf{6.20306}, \lambda_2 = \mathbf{3.20619}, \lambda_3 = \mathbf{2.71066}, \lambda_4 = \mathbf{1.88009}$$

The eigen values represent the variance of the data along the principal components. The first eigen value $\lambda_1 = 6.20306$ represents the variance of the data along the first principal component, the second eigen value $\lambda_2 = 3.20619$ represents the variance of the data along the second principal component, the third eigen value $\lambda_3 = 2.71066$ represents the variance of the data along the third principal component, and the fourth eigen value $\lambda_4 = 1.88009$ represents the variance of the data along the fourth principal component.

Part 2: Principal Component Analysis (PCA) (8 points)
Now that you have a grasp of the covariance matrix and its eigenvalues, you will apply PCA to a random vector

Principal Component Directions

(a) **(2 points)** Using the eigenvectors, describe the principal component directions. What do these directions represent in terms of variance in the data?

$$\text{Principal Component Directions} = \begin{bmatrix} v_1 & v_2 & v_3 & v_4 \end{bmatrix}$$

$$\text{Principal Component Directions} = \begin{bmatrix} 0.78907948 & 0.16328172 & 0.28804636 & 0.51741855 \\ 0.16466637 & -0.98478572 & 0.04206535 & 0.03622962 \\ 0.52644338 & 0.05882726 & 0.04585165 & -0.84693231 \\ 0.27036259 & 0.00869061 & -0.95559271 & 0.11692356 \end{bmatrix}$$

The principal component directions represent the directions along which the variance of the data is maximized. The directions are perpendicular to each other, which means that the principal components are independent of each other.

(b) **(2 points)** Explain the concept of orthogonality in PCA and why is it important?

Orthogonality in PCA means that the principal components are perpendicular to each other. This is important because it ensures that the principal components are independent of each other. This means that the variance of the data is maximized along the principal components.

Transformation of Random Vector
Let the eigenvector matrix be P and defined the transformed random vector Y by $Y = P^T X$

(a) **(2 points)** What is the covariance matrix of the vector Y?

$$\Sigma_Y = P^T \Sigma_x P$$

$$\Sigma_Y = \begin{bmatrix} 0.78907948 & 0.16466637 & 0.52644338 & 0.27036259 \\ 0.16328172 & -0.98478572 & 0.05882726 & 0.00869061 \\ 0.28804636 & 0.04206535 & 0.04585165 & -0.95559271 \\ 0.51741855 & 0.03622962 & -0.84693231 & 0.11692356 \end{bmatrix} \begin{bmatrix} 5 & 1.2 & 0.8 & 0.6 \\ 1.2 & 4 & 0.5 & 0.3 \\ 0.8 & 0.5 & 3 & 0.2 \\ 0.6 & 0.3 & 0.2 & 2 \end{bmatrix} \begin{bmatrix} 0.78907948 & 0.16328172 & 0.52644338 & 0.2703625 \\ 0.16466637 & -0.98478572 & 0.05882726 & 0.0086906 \\ 0.52644338 & 0.05882726 & 0.04585165 & -0.846932 \\ 0.27036259 & 0.00869061 & -0.95559271 & 0.1169235 \end{bmatrix}$$

$$\Sigma_Y = \begin{bmatrix} 6.20306 & 0 & 0 & 0 \\ 0 & 3.20619 & 0 & 0 \\ 0 & 0 & 2.71066 & 0 \\ 0 & 0 & 0 & 1.88009 \end{bmatrix}$$

(b) **(2 points)** How does this Transformation affect the correlation between the transformed features?
The transformation affects the correlation between the transformed features by making them uncorrelated. The covariance matrix of the transformed random vector Y is a diagonal matrix, which means that the transformed features are uncorrelated.

Part 3: Performing PCA by Hand on a Simple Dataset (8 points) consider a simple dataset represented by the following 2-dimensional random vector $Y = [Y_1, Y_2]^T$:

$$Y = \begin{bmatrix} 1.2 & 2.8 \\ 0.8 & 2.4 \\ 1.6 & 3.2 \\ 1.4 & 2.9 \end{bmatrix}$$

Mean Centering

(a) **(2 points)** Calculate the mean of the dataset for each feature Y_1 and Y_2

$$\text{Mean of } Y_1 = \frac{1.2 + 0.8 + 1.6 + 1.4}{4} = \mathbf{1.25}$$

$$\text{Mean of } Y_2 = \frac{2.8 + 2.4 + 3.2 + 2.9}{4} = \mathbf{2.825}$$

(b) **(2 points)** Subtract the mean from each feature to center the data

$$\text{Centered Data} = \begin{bmatrix} 1.2 - 1.25 & 2.8 - 2.825 \\ 0.8 - 1.25 & 2.4 - 2.825 \\ 1.6 - 1.25 & 3.2 - 2.825 \\ 1.4 - 1.25 & 2.9 - 2.825 \end{bmatrix} = \begin{bmatrix} -0.05 & -0.025 \\ -0.45 & -0.425 \\ 0.35 & 0.375 \\ 0.15 & 0.075 \end{bmatrix}$$

Covariance Matrix (2 points)

(a) Calculate the covariance matrix of the centered data

$$\text{Covariance Matrix} = \frac{1}{n-1}\text{Centered Data}^T\text{Centered Data}$$

$$\text{Covariance Matrix} = \frac{1}{4-1}\begin{bmatrix}-0.05 & -0.025\\ -0.45 & -0.425\\ 0.35 & 0.375\\ 0.15 & 0.075\end{bmatrix}^T\begin{bmatrix}-0.05 & -0.025\\ -0.45 & -0.425\\ 0.35 & 0.375\\ 0.15 & 0.075\end{bmatrix}$$

$$\text{Covariance Matrix} = \frac{1}{3}\begin{bmatrix}-0.05 & -0.45 & 0.35 & 0.15\\ -0.025 & -0.425 & 0.375 & 0.075\end{bmatrix}\begin{bmatrix}-0.05 & -0.025\\ -0.45 & -0.425\\ 0.35 & 0.375\\ 0.15 & 0.075\end{bmatrix}$$

$$\text{Covariance Matrix} = \frac{1}{3}\begin{bmatrix}0.35 & 0.335\\ 0.335 & 0.328\end{bmatrix}$$

$$\text{Covariance Matrix} = \begin{bmatrix}0.1167 & 0.1117\\ 0.1117 & 0.1093\end{bmatrix}$$

Eigenvalue decomposition (2 points)

(a) Manually compute the eigenvalues and eigenvectors of the covariance matrix
The Eigen values are obtained by solving the characteristic equation:

$$\text{Characteristic Equation: } \det(\Sigma - \lambda I) = 0$$

$$\begin{vmatrix}0.1167-\lambda & 0.1117\\ 0.1117 & 0.1093-\lambda\end{vmatrix} = 0$$

$$\lambda^2 - 0.226\lambda + 0.0003 = 0$$

$$\lambda_1 = 0.225, \lambda_2 = 0.0013$$

The Eigen vectors are obtained by solving the equation:
for eigen value $\lambda_1 = 0.225$

$$\begin{bmatrix}0.1167-0.225 & 0.1117\\ 0.1117 & 0.1093-0.225\end{bmatrix}\begin{bmatrix}v_1\\ v_2\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix}$$

$$\begin{bmatrix}-0.1083 & 0.1117\\ 0.1117 & -0.1157\end{bmatrix}\begin{bmatrix}v_1\\ v_2\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix}$$

$$R_1 = \frac{1}{-0.1083}R_1 = \begin{bmatrix}1 & -1.031 & | & 0\\ 0.1117 & -0.1157 & | & 0\end{bmatrix}$$

$$R_2 = R_2 - 0.1117R_1 = \begin{bmatrix}1 & -1.031 & | & 0\\ 0 & -0.0005 & | & 0\end{bmatrix}$$

$$R_2 = \frac{1}{-0.0005}R_2 = \begin{bmatrix}1 & -1.031 & | & 0\\ 0 & 1 & | & 0\end{bmatrix}$$

$$v_1 = 1.031v_2$$

$$v_1 = \begin{bmatrix}1.031\\ 1\end{bmatrix}$$

for eigen value $\lambda_2 = 0.0013$

$$\begin{bmatrix}0.1167-0.0013 & 0.1117\\ 0.1117 & 0.1093-0.0013\end{bmatrix}\begin{bmatrix}v_1\\ v_2\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix}$$

$$\begin{bmatrix}0.1154 & 0.1117\\ 0.1117 & 0.1080\end{bmatrix}\begin{bmatrix}v_1\\ v_2\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix}$$

$$R_1 = \frac{1}{0.1154}R_1 = \begin{bmatrix}1 & 0.968 & | & 0\\ 0.1117 & 0.1080 & | & 0\end{bmatrix}$$

$$R_2 = R_2 - 0.1117R_1 = \begin{bmatrix}1 & 0.968 & | & 0\\ 0 & -0.0001256 & | & 0\end{bmatrix}$$

$$R_2 = \frac{1}{-0.0001256}R_2 = \begin{bmatrix}1 & 0.968 & | & 0\\ 0 & 1 & | & 0\end{bmatrix}$$

$$v_1 = -0.968v_2$$

$$v_1 = \begin{bmatrix}-0.968\\ 1\end{bmatrix}$$

so the eigen vectors are as follows:

$$\lambda_1 = 0.225, v_1 = \begin{bmatrix}1.031\\ 1\end{bmatrix}$$

$$\lambda_2 = 0.0013, v_2 = \begin{bmatrix}-0.968\\ 1\end{bmatrix}$$

Project the data (2 points)

(a) Using the principal component corresponding to the largest eigenvalue, project the original data onto the principal component axis

$$\text{Projection} = \text{Centered Data} \times v_1$$

$$\text{Projection} = \begin{bmatrix} -0.05 & -0.025 \\ -0.45 & -0.425 \\ 0.35 & 0.375 \\ 0.15 & 0.075 \end{bmatrix} \begin{bmatrix} 1.031 \\ 1 \end{bmatrix}$$

$$\text{Projection} = \begin{bmatrix} -0.05 \times 1.031 + -0.025 \times 1 \\ -0.45 \times 1.031 + -0.425 \times 1 \\ 0.35 \times 1.031 + 0.375 \times 1 \\ 0.15 \times 1.031 + 0.075 \times 1 \end{bmatrix}$$

$$\text{Projection} = \begin{bmatrix} -0.07655 \\ -0.88895 \\ 0.73585 \\ 0.22965 \end{bmatrix}$$

(b) Show the final transformed data in 1D (along the principal component axis)

$$\text{Final Transformed Data} = \begin{bmatrix} -0.07655 \\ -0.88895 \\ 0.73585 \\ 0.22965 \end{bmatrix}$$

PCA in Practice with a Large Dataset (10 points)

You are now provided with a dataset consisting of 500 users, where each user has four features: Usage time, Interactions, Activity type 1, Activity type 2. You will apply PCA using python to reduce the dimensionality

[**in the notebook**]

Intepretation & Business Insights (20 points)

Part 1: Feature Intepretation in PCA:

(a) **(5 points)** Based on the principal components directions, explain which features (original dimensions) contribute most to the first and second principal components.
The first principal component is most influenced by the feature Usage time, while the second principal component is most influenced by Activity type 1. This means that Usage time has the most significant impact on the variance of the data along the first principal component, while Activity type 1 has the most significant impact on the variance of the data along the second principal component.

(b) **(5 points)** How would you explain the reduced features to a non-technical team in terms of the user behavior patterns?
The reduced features represent the most important aspects of user behavior patterns. The first principal component captures the overall usage time of the users, while the second principal component captures the type of activities users engage in. By focusing on these two principal components, we can understand the key factors that drive user behavior and make informed decisions based on these patterns.

Part 2: Using PCA for Future Decision-Making:

(a) (5 points) How can the startup use the reduced-dimensional data for faster processing and improved decision-making?
The startup can use the reduced-dimensional data to streamline its data processing and improve decision-making in several ways. By focusing on the most important features captured by the principal components, the startup can reduce the complexity of its data analysis and speed up processing times. This will allow the startup to make faster and more accurate decisions based on the key patterns in the data. Additionally, the reduced-dimensional data can help the startup identify trends and insights that may not be apparent in the original high-dimensional data, enabling more effective strategic planning and resource allocation.

(b) (5 points) What are the potential risks of reducing the dimensionality in this manner? Could important information be lost, and how would you mitigate this?
Reducing the dimensionality of the data using PCA can lead to the loss of important information that may be present in the original high-dimensional data. By focusing on the most significant features captured by the principal components, the startup may overlook subtle patterns or relationships that could be valuable for decision-making. To mitigate this risk, the startup should carefully evaluate the trade-offs between dimensionality reduction and information loss. It should also consider using other techniques, such as feature selection or feature engineering, to retain critical information while reducing the complexity of the data. Additionally, the startup should regularly validate the results of the dimensionality reduction process to ensure that important insights are not overlooked.

Question 2: SUM OF RANDOM VARIABLES, CENTRAL lIMIT THEOREM & PROBABILITY BOUNDS (5O marks)

Reading: The sum of random variables is a fundamental concept in probability and statistics, shedding light on the behavior of combined outcomes under uncertainty. When adding two random variables, X and Y $(i.e., Z = X + Y)$, we analyze the distribution of Z. For independent variables, the distribution of Z can be derived by convolving the individual distributions.
For example, if both X and Y are normally distributed with means μ_X and μ_Y and variances σ_X^2 and σ_Y^2, then Z will also be normally distributed, with mean $\mu_Z = \mu_X + \mu_Y$ and variance $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$ which is beneficial for statistical modeling.
When X and Y are not independent, we must include covariance, which accounts for how the variables change together, in the variance of Z: $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 + 2\text{Cov}(X, Y)$ + 2Cov(X, Y).
This concept applies to any number of random variables. For independent variables $X_1, X_2, ..., X_n$, the sum $Z = X_1 + X_2 + ... + X_n$ is analyzed similarly. The Central Limit Theorem indicates that as the number of independent variables increases, their standardized sum approaches a normal distribution. Understanding the sum of random variables is essential in finance, insurance, and natural sciences for risk assessment, forecasts, and decision-making, revealing insights into complex systems. Probability bounds are crucial in statistics, quantifying the likelihood of events within specified limits, particularly in finance and engineering.
Markov's inequality estimates the probability that a non-negative random variable X exceeds a certain value a, stating that for any $a > 0$, the probability that $X \geq a$ is at most the expected value of X divided by a. This demonstrates that limited information about X's distribution can provide useful probability estimates.
Chebyshev's inequality extends Markov's by considering variance. It states that for any random variable X with mean μ and finite variance σ^2, the probability that X deviates from its mean by more than k standard deviations is at most 1 k2 . This finding is fundamental for statistical inference. Hoeffding's inequality offers bounds for sums of independent random variables, ensuring exponential decay in tail probabilities, which is especially valuable in large sample scenarios to keep observed averages close to expected values.
Overall, these probability bounds enhance decision-making and deepen our understanding of stochastic processes, facilitating robust conclusions across various fields.

Part 1: Mobile Network Data Analysis (10 points)
In a study conducted by a telecommunication company in Rwanda, mobile network Calls are classified as either voice (V) when someone is speaking or data (D) when there is a modem or fax transmission. Based on observed data, the probabilities are:

$$P(V) = 0.6 \text{ (60\% voice calls)}$$

$$P(D) = 0.4 \text{(40\% data calls)}$$

Assume data calls and voice calls occur independently of each other, and let the random variable K_n represent the number of data calls in a collection of n calls.

1. **(2 points)** What is the $E[K_{100}]$, the expected number of data calls in a set of 100 calls?

$$E[K_{100}] = n \times P[D] = 100 \times 0.4 = \mathbf{40}$$

2. **(2 points)** What is $\sigma_{k_{100}}$, the standard deviation of the number of voice calls in a set of 100 calls?

$$\sigma_{k_{100}} = \sqrt{n \times P[D] \times P[V]} = \sqrt{100 \times 0.4 \times 0.6} = \mathbf{4.9}$$

3. **(2 points)** Using Central Limit Theorem, estimate $P[K_{100} \geq 18]$, the probability of at least 18 data calls in a set of 100 calls
Find the probability of at least 18 data calls in a set of 100 calls:

$$P[K_{100} \geq 18] = 1 - P[K_{100} \leq 17]$$

The Z-score for 17 data calls is:

$$Z = \frac{17 - 40}{4.9} = -4.69387755$$

This means that the probability of at least 18 data calls is:

$$P[K_{100} \geq 18] = 1 - P[Z \leq -4.9] = 1 - 0 = \mathbf{1}$$

4. **(2 points)** Using the CLT, estimate $P[16 \leq K_{100} \leq 24]$, the probability of between 16 and 24 data calls in a set of 100 calls

$$P[16 \leq K_{100} \leq 24] = P[K_{100} \leq 24] - P[K_{100} \leq 16]$$

$$P[K_{100} \leq 24] = P\left[\frac{K_{100} - E[K_{100}]}{\sigma_{K_{100}}} \leq \frac{24 - 40}{4.9}\right] = P[Z \leq -3.27]$$

$$P[Z \leq -3.27] = 0.0005$$

$$P[K_{100} \leq 16] = P\left[\frac{K_{100} - E[K_{100}]}{\sigma_{K_{100}}} \leq \frac{16 - 40}{4.9}\right] = P[Z \leq -4.9]$$

$$P[Z \leq -4.9] = 0$$

$$P[16 \leq K_{100} \leq 24] = 0.0005 - 0 = \mathbf{0.0005}$$

5. **(2 points)** Based on your calculations, what can you infer about the likelyhood of high data traffic during a given period? How might this information help a telecom optimize their resources for voice and data services?

The likelihood of high data traffic during a given period is very low. This information can help a telecom optimize their resources for voice and data services by ensuring that they have enough capacity to handle voice calls, which are more likely to occur. They can also allocate resources to handle data calls, but they do not need to allocate as many resources since data calls are less likely to occur.

Part 2: Chernoff Bound & Gaussian Random Variables (4 points)
Use the Chernoff bound to show that for a Gaussian (Normal) random variable X with mean μ and standard deviation σ, the probability that X exceeds a certain threshold c can be bounded by:

$$P[X \geq c] \leq e^{-\frac{(c-\mu)^2}{2\sigma^2}}$$

proof:

$$P[X \geq c] = P[e^{tX} \geq e^{tc}]$$
$$P[X \geq c] = P[e^{tX} \geq e^{tc}] \leq \frac{E[e^{tX}]}{e^{tc}}$$
$$E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$
$$E[e^{tX}] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{tx-\frac{(x-\mu)^2}{2\sigma^2}} dx$$
$$E[e^{tX}] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{tx-\frac{x^2-2x\mu+\mu^2}{2\sigma^2}} dx$$

but the integral is the Gaussian integral:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{tx-\frac{x^2-2x\mu+\mu^2}{2\sigma^2}} dx = e^{\frac{t^2\sigma^2}{2}+t\mu}$$

solving for t:

$$e^{\frac{t^2\sigma^2}{2}+t\mu} = e^{tc}$$
$$\frac{t^2\sigma^2}{2} + t\mu = tc$$
$$\frac{t\sigma^2}{2} + \mu = c$$
$$t = \frac{2(c-\mu)}{\sigma^2}$$

substituting back into the Chernoff bound:

$$P[X \geq c] \leq \frac{e^{\frac{2(c-\mu)}{\sigma^2}X}}{e^{tc}}$$
$$P[X \geq c] \leq e^{-\frac{(c-\mu)^2}{2\sigma^2}}$$

Given this result, how would you use it to provide a worst-case scenario estimate in a real-world context, such as predicting an extreme event like an abnormally high network traffic spike or stock price surge?

Part 3: Soccer Tournament Performance (11 points)
Manchester United is competing in a knockout-style tournament, where each game can result in a win, loss, or tie. For every win, they earn 3 points, for every tie 1 point, and for a loss 0 points. The outcome of each game is independent of the others, and each game result is equally likely (win, loss, or tie). Let X_i be the number of points earned in game i, and Y represent the total number of points earned over the course of the tournament.

1. **(3 points)** Derive the moment generating function (MGF) of ϕ_Y(s)

$$\phi_Y(s) = E[e^{sY}]$$

where $Y = X_1 + X_2 + X_3 + \cdots + X_n$, the sum of the random variables and s is the parameter of the MGF
random variables X_i are independent and identically distributed, so the MGF of the sum of the random variables is the product of the MGFs of the individual random variables
A random variable X_i can take on the values 0, 1, 3 with probabilities $\frac{1}{3}$ each

$$\phi_{X_i}(s) = E[e^{sX_i}] = \frac{1}{3}e^0 + \frac{1}{3}e^s + \frac{1}{3}e^{3s}$$
$$\phi_{X_i}(s) = \frac{1}{3} + \frac{1}{3}e^s + \frac{1}{3}e^{3s}$$
$$\phi_{X_i}(s) = \frac{1}{3}(1 + e^s + e^{3s})$$
$$\phi_Y(s) = \phi_{X_1}(s) \times \phi_{X_2}(s) \times \phi_{X_3}(s) \times \phi_{X_4}(s) \cdots \times \phi_{X_n}(s)$$
$$\phi_Y(s) = \left(\frac{1}{3}(1 + e^s + e^{3s})\right)^n$$

2. **(5 points)** Find E[Y] and Var[Y], the expected total points and variance.
Our moment generating function is:

$$\phi_Y(s) = \left(\frac{1}{3}(1 + e^s + e^{3s})\right)^n$$

to get the expected value, we differentiate the MGF with respect to s and evaluate at s = 0:

$$E[Y] = \phi_Y'(s) = n\left(\frac{1}{3}(1 + e^s + e^{3s})\right)^{n-1} \times \frac{d}{ds}\left(\frac{1}{3}(1 + e^s + e^{3s})\right)$$

$$E[Y] = n\left(\frac{1}{3}(1+e^s+e^{3s})\right)^{n-1} \times \frac{1}{3}(e^s+3e^{3s})$$

set s = 0:

$$E[Y] = n\left(\frac{1}{3}(1+1+1)\right)^{n-1} \times \frac{1}{3}(1+3)$$

$$E[Y] = n\left(\frac{3}{3}\right)^{n-1} \times \frac{4}{3}$$

so the expected value of Y is:

$$E[Y] = \frac{\mathbf{4n}}{\mathbf{3}}$$

The variance is obtained by differentiating the MGF twice with respect to s and evaluating at s = 0:

$$Var[Y] = \phi''_Y(s) = n(n-1)\left(\frac{1}{3}(1+e^s+e^{3s})\right)^{n-2} \times \frac{d^2}{ds^2}\left(\frac{1}{3}(1+e^s+e^{3s})\right)$$

$$Var[Y] = n(n-1)\left(\frac{1}{3}(1+e^s+e^{3s})\right)^{n-2} \times \frac{d}{ds}\left(\frac{1}{3}(e^s+3e^{3s})\right)$$

$$Var[Y] = n(n-1)\left(\frac{1}{3}(1+e^s+e^{3s})\right)^{n-2} \times \frac{1}{3}(e^s+9e^{3s})$$

set s = 0:

$$Var[Y] = n(n-1)\left(\frac{1}{3}(1+1+1)\right)^{n-2} \times \frac{1}{3}(1+9)$$

$$Var[Y] = n(n-1)\left(\frac{3}{3}\right)^{n-2} \times \frac{10}{3}$$

so the variance of Y is:

$$Var[Y] = \frac{\mathbf{10n(n-1)}}{\mathbf{9}}$$

3. **(3 points)** Based on your calculations, what can you infer about the likely performance of Manchester United over the course of multiple tournaments? How might the expected points impact their overall ranking or their chances of advancing in the competition?

   Based on the calculations, we can infer that the expected performance of Manchester United over the course of multiple tournaments is to earn an average of $\frac{4n}{3}$ points with a variance of $\frac{10n(n-1)}{9}$. This information can help predict their overall ranking in the competition and their chances of advancing. Teams with higher expected points and low variance are more likely to rank higher and advance in the competition, while teams with lower expected points are less likely to do so.

Part 4: Course Enrollment and Resource Planning (6 points)
The number of students enrolling in a popular data science course is modelled as a poisson random variable with a mean of 100 students. The Professor has decided that if 120 students enroll, he will split the class into two sectioons, otherwise , he will teach all the students in a single Section

1. **(3 points)** What is the probability that the professor will need to teach two sections?
   using Markov's inequality:
   The markov inequality is:

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

where a is the threshold value and E[X] is the expected value of Xg

$$P[X \geq 120] \leq \frac{E[X]}{120}$$

$$P[X \geq 120] \leq \frac{100}{120}$$

$$P[X \geq 120] \leq 0.833$$

$$P[X \geq 120] = \mathbf{0.833}$$

2. **(3 points)** Based on Probability, what recommendations would you make regarding resource planning for future courses? Should the professor prepare for two sections or allocate resources differently based on expected enrollments?

   Based on the probability that the professor will need to teach two sections, it is recommended that the professor prepare for two sections. Since the probability is 0.833, which is greater than 0.5, it is more likely that 120 students will enroll. By preparing for two sections, the professor can ensure that there are enough resources to accommodate the students and provide a better learning experience.

Part 5: Comparison of Markov, Chebyshev and Chernoff Inequalities (19 points) consider a Gaussian Random Variable X with mean $\mu = 0$ and variance $\sigma = 1$. We are interested in comparing how well three probability probability bounds - **Markov, Chebyshev and Chernoff** - estimate the probability that X exceeds a certain threshold c. i.e $P[X \geq c]$

1. **Markov Inequality**
   For a non-negative random variable X, Markov's inequality states that for any $c > 0$, the probability that $X \geq a$ is at most the expected value of X divided by a. This provides a simple bound on the tail probability of a random variable.

   $$P[X \geq c] \leq \frac{E[X]}{c}$$

   For this comparison, we will apply Markov's inequality to the positive part of the Gaussian random variable X, considering $X^+ = \max(X, 0)$.

2. **Chebyshev Inequality**
   Chebyshev's inequality, applicable to any random variable with finite variance, states that for any random variable X with mean μ and finite variance σ^2, the probability that X deviates from its mean by more than k standard deviations is at most 1/k2. This provides a more refined bound on the tail probability of a random variable.

   $$P[|X - \mu| \geq c] \leq \frac{\sigma^2}{c^2}$$

   We will apply this inequality to the Gaussian random variable

3. **Chernoff bound**
   For a Gaussian random Variable X, the Chernoff bound provides a tighter bound for tail probabilities

   $$P[X \geq c] \leq \exp(-\frac{(c-\mu)^2}{2\sigma^2})$$

   The bound is especially for normally distributed data.

4. **Comparison**

   [in the notebook]

Question 3: ESTIMATION & HYPOTHESIS TESTING, CONFIDENCE INTERVALS (130 marks)

Reading: Estimation and hypothesis Testing provide a framework for making inference about populations based on sample data. **Estimation** is the process of inferring population parameters (such as means or proportions) from sample statistics. There are two main types of estimations:

Point Estimation

Provides a single value as the estimate of a population parameter.
Example: The sample mean $\bar{X}$ is a point estimate of the population mean μ.
Formula:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Interval Estimation

Offers a range of values within which the population parameter is likely to fall. These values are in arange called confidence interval, believed to contain the parameter based with a specific level of confindence:
Example: A 95% confidence interval for the population mean μ is given by $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$
Formula:

$$\text{Confidence Interval} = \bar{X} \pm z \times \frac{\sigma}{\sqrt{n}}$$

where Z is the z-score corresponding to the desired confidence level.

Hypothesis Testing

Hypothesis testing is a statistical method used to make inferences about population parameters based on sample data. You use it to make probabilistic decisions about the population based on sample data. The process involves:

Step 1: Formulate Hypotheses

The null hypothesis (H_0) is the default assumption that there is no effect or no difference.
The alternative hypothesis (H_1 or H_a) is the claim that there is an effect or difference. The claim to be tested

Step 2: Set the Significance Level

The significance level (α) is the probability of rejecting the null hypothesis when it is true. Common values are 0.05 or 0.01. This is the threshold of rejecting the null hypothesis.

Step 3: Calculate the Test Statistic

The test statistic is a numerical summary of the data that measures the difference between the sample data and the null hypothesis. It is used to determine whether the null hypothesis should be rejected.

$$\text{Test Statistic} = \frac{\text{Observed Value} - \text{Expected Value}}{\text{Standard Error}}$$

$$\text{Standard Error} = \frac{\text{Standard Deviation}}{\sqrt{n}}$$

$$\text{Observed Value} = \text{Sample Mean}$$

$$\text{Expected Value} = \text{Population Mean}$$

$$\text{Standard Deviation} = \text{Population Standard Deviation}$$

$$\text{n} = \text{Sample Size}$$

the final formula is:

$$\text{Test Statistic} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Step 4: Make a Decision

Compare the test statistic to the critical value. If the test statistic falls in the rejection region, reject the null hypothesis. If it falls in the non-rejection region, do not reject the null hypothesis.

$$\text{Rejection Region} = \text{Critical Value}$$

$$\text{if } Z \geq \text{Critical Value, reject } H_0$$

where the critical value is determined by the significance level and the type of test (one-tailed or two-tailed).

Step 5: outcome:

Based on the decision, you can either reject or fail to reject the null hypothesis. This decision provides insight into the population parameter and the relationship between the sample and the population.
if the null hypothesis is rejected, it suggests strong evidence in favor of the alternative hypothesis
if not rejected, there isn't sufficient evidence to support the alternative hypothesis

Central Limit Theorem

It plays a crucial role in both estimation and hypothesis testing stating that the sample mean $\bar(X)$ will approach a normal distribution as the sample size(n), increases, regardless of the population's distribution
This Theorem provides normal approximation techniques in practice, facilitating easier calculations and Intepretation in estimation and hypothesis testing..

Part 1: Point Estimation: Estimating the Average Battery Life (20 marks)

A company is producing a new smartphone model and wants to estimate the average Battery life. From a sample of 20 smartphones, the following battery life data (in hours) is collected:

$$\text{Battery Life} = [8.2, 8.5, 8.9, 9.0, 7.8, 8.6, 8.4, 8.1, 9.1, 8.7, 9.2, 8.8, 8.3, 9.3, 8.0, 8.9, 8.4, 8.6, 8.7, 8.2]$$

1. **(2 points)** Calculate the sample mean and sample variance.

$$\text{Sample Mean} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\text{Sample Mean} = \frac{1}{20}(8.2 + 8.5 + 8.9 + 9.0 + 7.8 + 8.6 + 8.4 + 8.1 + 9.1 + 8.7 + 9.2 + 8.8 + 8.3 + 9.3 + 8.0 + 8.9 + 8.4 + 8.6 + 8.7 + 8.2)$$

$$\text{Sample Mean} = \frac{1}{20}(171.7) = \mathbf{8.585}$$

sample variance:

$$\text{Sample Variance} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\text{Sample Variance} = \frac{1}{19}((8.2{-}8.585)^2{+}(8.5{-}8.585)^2{+}(8.9{-}8.585)^2{+}(9.0{-}8.585)^2{+}(7.8{-}8.585)^2{+}(8.6{-}8.585)^2{+}(8.4{-}8.585)^2{+}(8.1{-}8.585)^2$$
$$+(9.1{-}8.585)^2{+}(8.7{-}8.585)^2{+}(9.2{-}8.585)^2{+}(8.8{-}8.585)^2{+}(8.3{-}8.585)^2{+}(9.3{-}8.585)^2{+}(8.0{-}8.585)^2{+}(8.9{-}8.585)^2{+}(8.4{-}8.585)^2{+}(8.6{-}8.585)^2$$
$$+(8.7 - 8.585)^2 + (8.2 - 8.585)^2)$$

$$\text{Sample Variance} = \frac{1}{19}(3.0725425) = \mathbf{0.1617}$$

2. **(3 points)** What does the sample mean estimate in this case, and determine if the sample mean is unbiased estimator of the population μ. Explain your resonings.
The sample mean estimates the average battery life of the smartphones in the sample. The sample mean is an unbiased estimator of the population mean μ if the expected value of the sample mean is equal to the population mean. In this case, the sample mean is an unbiased estimator of the population mean μ because the expected value of the sample mean is equal to the population mean.

3. **(3 points)** Compute the mean square error (MSE) of the sample mean and variance.

$$\text{MSE} = \text{Bias}(\bar{X})^2 + \text{Variance}(\bar{X})$$

$$\text{Bias}(\bar{X}) = E[\bar{X}] - \mu$$

$$\text{Variance}(\bar{X}) = 0.1617$$

$$\text{MSE} = (8.585 - 8.585)^2 + 0.1617 = \mathbf{0.1617}$$

4. (2 points)Explain how the error of the mean and the sample variance might change if the sample size were smaller or larger.

   If the sample size were smaller, the error of the mean and the sample variance would likely be larger because the sample mean and sample variance would be less representative of the population. If the sample size were larger, the error of the mean and the sample variance would likely be smaller because the sample mean and sample variance would be more representative of the population.

5. (2 points) What effect do outliers have on the sample mean and variance?

   If there are outliers in the sample, they can skew the sample mean and variance, making them less representative of the population. Outliers can also increase the variance of the sample, making it less accurate as an estimate of the population variance.

6. (8 points) Dynamically simulate different sample sizes $n \in [5, 1000, \textit{step} = 5]$ and generate samples from a normal distribution with a known population mean $\mu = 8.5$ and standard deviation $\sigma = 0.5$. Vary n and compare how the sample mean and variance behave as the sample size changes

Part 2: Confidence Intervals; Estimating the True Mean Height (16 marks)
You are studying the average height of adult males in a city. A random sample of 30 men yields a sample mean height of **176 cm** and a standard deviation of **7 cm**.

1. **(2 points)** Calculate a 95% confidence interval for the population mean height, assuming that the population standard deviation is unkown

$$\text{Confidence Interval} = \bar{X} \pm z \times \frac{\sigma}{\sqrt{n}}$$

$$\text{Confidence Interval} = 176 \pm 1.96 \times \frac{7}{\sqrt{30}}$$

$$\text{Confidence Interval} = 176 \pm 1.96 \times \frac{7}{5.48}$$

$$\text{Confidence Interval} = 176 \pm 1.96 \times 1.28$$

$$\text{Confidence Interval} = 176 \pm 2.51$$

$$\text{Confidence Interval} = [173.49, 178.51]$$

2. **(2 points)** If the population standard deviation was known to be 7 cm, how would this change the confidence interval?

$$\text{Confidence Interval} = \bar{X} \pm z \times \frac{\sigma}{\sqrt{n}}$$

$$\text{Confidence Interval} = 176 \pm 1.96 \times \frac{7}{\sqrt{30}}$$

$$\text{Confidence Interval} = 176 \pm 1.96 \times \frac{7}{5.48}$$

$$\text{Confidence Interval} = 176 \pm 1.96 \times 1.28$$

$$\text{Confidence Interval} = 176 \pm 2.51$$

$$\text{Confidence Interval} = [173.49, 178.51]$$

The confidence interval would be the same because the population standard deviation does not affect the calculation of the confidence interval when it is known.

3. **(2 points)** What does the confidence interval mean in practical terms?

The confidence interval means that we are 95% confident that the true mean height of the men is between 173.49 cm and 178.51 cm. This interval provides a range of values within which the population mean height is likely to fall with a 95% confidence level.

4. **(8 points)** Simulate the construction of confidence intervals for different sample sizes, and confidence levels, Vary $n \in [10, 500, \text{step} = 10]$ and $\alpha \in [0.01, 0.2, \text{step} = 0.01]$ dynamically and observe how the confidence interval width changes.

[in the notebook]

Part 3: Hypothesis Testing; Comparing Two Webpage Designs (A/B Testing) (19 points)
You are running an A/B test for two different webpage designs to see which one generates more clicks. Out of 600 visitors, 240 clicked on Version A, and 290 clicked on Version B.

1. **(2 points)** Calculate the proportions of clicks for each version

$$\text{Proportion of Clicks for Version A} = \frac{240}{600} = \mathbf{0.4}$$

$$\text{Proportion of Clicks for Version B} = \frac{290}{600} = \mathbf{0.4833}$$

2. **(2 points)** Perform a hypothesis test to determine whether there is a significant difference between the click-through rates (CTR) of the two versions. Assume $\alpha = 0.05$ and compute the z-statistic by hand

$$\text{Null Hypothesis}(H_0) : \text{There is no significant difference between the click-through rates of the two versions}$$

$$\text{Alternative Hypothesis}(H_1) : \text{There is a significant difference between the click-through rates of the two versions}$$

$$p = \frac{240 + 290}{600 + 600} = \frac{530}{1200} = 0.4417$$

$$\text{Standard Error} = \sqrt{0.4417(1 - 0.4417)(\frac{1}{600} + \frac{1}{600})} = 0.02867$$

$$\text{Z-Statistic} = \frac{0.4 - 0.4833}{0.02867} = -2.905$$

$$\text{Critical Value} = -1.96$$

$$\text{if } Z \leq -1.96, \text{ reject } H_0$$

$$-2.905 \leq -1.96, \text{ reject } H_0$$

$$\text{Reject } H_0$$

3. **(2 points)** What does the p-value mean in the context of this A/B test?

$$\text{P-Value} = \mathbf{0.0036}$$

In this case, the p-value is less than the significance level of 0.05, indicating that the observed difference in click-through rates is statistically significant.

4. **(3 points)** How would you interpret the results to your marketing team?

The results of the A/B test indicate that there is a significant difference between the click-through rates of the two versions of the webpage. Version B has a higher click-through rate than Version A, suggesting that it may be more effective in generating clicks from visitors. This information can be used by the marketing team to make data-driven decisions about which webpage design to use in the future.

5. **(6 points)** Use a dynamic simulation to vary the sample sizes $n \in [100, 1000, \text{step} = 50]$ and click-through rates $c \in [50, 600, \text{step} = 10]$ for Versions A and B. Observe how the p-value and hypothesis test decision change as you modify these parameters.

[in the notebook]

Part 4: True Positives and False Positives in Medical Testing (23 points)
A medical test for a rare disease is conducted on 1000 people. The test correctly identifies 80 true positives and 900 true negatives. However, it also produces 10 false positives and 10 false negatives.

1. (5 points) Construct a confusion matrix from the data provided.

	Actual Positive	Actual Negative
Predicted Positive	80	10
Predicted Negative	10	900

2. (4 points) Compute the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the test.
   sensitivity - recall:
$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} = \frac{80}{80+10} = \mathbf{0.8889}$$
   specificity - precision:
$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} = \frac{900}{900+10} = \mathbf{0.9889}$$
   positive predictive value:
$$\text{PPV} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} = \frac{80}{80+10} = \mathbf{0.8889}$$
   negative predictive value:
$$\text{NPV} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Negatives}} = \frac{900}{900+10} = \mathbf{0.9889}$$

3. (4 points) What do these values indicate about the reliability of the medical test?

   The sensitivity, specificity, positive predictive value, and negative predictive value of the test are all high, indicating that the test is reliable in identifying both positive and negative cases of the disease. The high sensitivity and specificity values suggest that the test has a low rate of false positives and false negatives, making it an effective diagnostic tool.

4. (3 points) Suppose the population size increases to 5000, with the same sensitivity and specificity. How would this affect the number of true and false positives?
$$\text{True Positives} = 0.8889 \times 5000 = \mathbf{4444.5}$$
$$\text{False Positives} = 0.1111 \times 5000 = \mathbf{555.5}$$

5. **(7 points)** Simulate different population sizes and rates of true positives and false positives using sliders. Visualize how the sensitivity, specificity, PPV, and NPV change as the sample size and the number of positive cases fluctuate.

   [**in the notebook**]

Part 5: MLE/MAP (14 marks)

In this problem, we have a coin for which we wish to determine the bias (i.e., is it a fair coin that shows heads 50% of the time, or is it biased with some other probability of showing heads?). We will consider eight possible hypotheses (each representing the probability that heads is flipped):

Hypothesis	Probability of Heads
H1	0%
H2	15%
H3	30%
H4	45%
H5	60%
H6	75%
H7	90%
H8	100%

Table 1: Coin toss observed under various hypotheses

1. **(4 points)** Given the first five coin flips: [1, 0, 1, 1, 1], calculate the likelihood of these observation under each hypothesis.
$$\text{Likelihood} = p^{n_{\text{heads}}} \times (1-p)^{n_{\text{tails}}}$$
$$\text{Likelihood} = p^4 \times (1-p)^1$$
   For H1:
$$\text{Likelihood} = 0 \times (1-0)^1 = \mathbf{0}$$
   For H2:
$$\text{Likelihood} = 0.15^4 \times (1-0.15)^1 = \mathbf{0.00043}$$
   For H3:
$$\text{Likelihood} = 0.3^4 \times (1-0.3)^1 = \mathbf{0.000567}$$
   For H4:
$$\text{Likelihood} = 0.45^4 \times (1-0.45)^1 = \mathbf{0.02255}$$
   For H5:
$$\text{Likelihood} = 0.6^4 \times (1-0.6)^1 = \mathbf{0.05184}$$
   For H6:
$$\text{Likelihood} = 0.75^4 \times (1-0.75)^1 = \mathbf{0.07910}$$
   For H7:
$$\text{Likelihood} = 0.9^4 \times (1-0.9)^1 = \mathbf{0.06561}$$
   For H8:
$$\text{Likelihood} = 1^4 \times (1-1)^1 = \mathbf{0}$$

2. **(10 points)** Given the coin flip results: [1, 0, 1, 1, 1, 0, 0, 1, 1, 1], perform an MAP and MLE experiment to determine the probability of each hypothesis with respect to the number of coin flips

   (a) Generate a plot showing the posterior probability of each hypothesis with respect to the number of observations

$$\text{Posterior Probability} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

$$\text{Posterior Probability} = \frac{\text{Likelihood} \times \text{Prior}}{\sum_{i=1}^{8} \text{Likelihood}_i \times \text{Prior}_i}$$

[in the notebook]

   (b) Generate a plot showing the probability that the next coin flip is heads with respect to the number of observations

[in the notebook]

   (c) What is the most likely hypothesis after all observations are made?
   The most likely hypothesis after all observations are made is the one with the highest posterior probability. In this case, the hypothesis with the highest posterior probability is H6 (75% probability of heads).

Part 6: Students' Exam Performance (23 points)

You want to determine whether students who use cheat sheets perform better on average relative to those who do not. To test this hypothesis, you have collected scores from two groups of students: one group that used cheat sheets and one that did not. Consider both sample sizes of 45 and 55 observations, respectively, for those that use cheat sheets and those who do not. If the mean scores of those who use cheat sheets in the exam is 88 and those who do not is 85 with standard deviations of 3 and 2 respectively, assume the samples are normally distributed.

1. **(2 points)** Given the sample sizes, means, and standard deviations for both groups, calculate the pooled standard deviation (S_p) and the standard error (SE)
   Pooled Standard Deviation is the weighted average of the standard deviations of the two samples, taking into account the sample sizes and variances:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

   where:

$$S_1 = 3, S_2 = 2, n_1 = 45, n_2 = 55$$

$$S_p = \sqrt{\frac{(45 - 1)3^2 + (55 - 1)2^2}{45 + 55 - 2}}$$

$$S_p = \sqrt{\frac{44 \times 9 + 54 \times 4}{98}}$$

$$S_p = \sqrt{\frac{396 + 216}{98}}$$

$$S_p = \sqrt{\frac{612}{98}}$$

$$S_p = \sqrt{6.24} = \mathbf{2.49898}$$

   Standard Error:

$$SE = s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$SE = 2.49898 \times \sqrt{\frac{1}{45} + \frac{1}{55}}$$

$$SE = 2.49898 \times \sqrt{0.02222 + 0.01818}$$

$$SE = 2.49898 \times \sqrt{0.0404}$$

$$SE = 2.49898 \times 0.201$$

$$SE = \mathbf{0.522}$$

2. **(2 points)** Calculate the z-score for the difference in sample means and the critical score for a one-tailed z-test at a significance of 5%

$$\text{Z-Score} = \frac{\bar{X}_1 - \bar{X}_2}{SE}$$

$$\text{Z-Score} = \frac{88 - 85}{0.522}$$

$$\text{Z-Score} = \frac{3}{0.522} = \mathbf{5.74}$$

3. **(2 points)** Based on your calculated z-score and the critical value, do you reject or fail to reject the null hypothesis? What does this imply about the performance of students using cheat sheets?

   Since the calculated z-score of 5.74 is greater than the critical value of 1.645 at a significance level of 5%, we reject the null hypothesis. This implies that students using cheat sheets perform significantly better on average compared to those who do not.

4. **(2 points)** Calculate the p-value for your test statistic. What does the p-value indicate about the significance of your results?
p-value:
You read the p-value from the z-table

$$\text{P-Value} = \mathbf{0.0001}$$

The p-value of 0.0001 indicates that the observed difference in performance between the two groups is highly statistically significant. The p-value is less than the significance level of 0.05, providing strong evidence to reject the null hypothesis.

5. **(15 points)** Simulate an examination process to determine whether students using cheat sheets perform better on average compared to those who do not. Perform a z-score test to verify this claim, simulate the process, and represent the results using interactive plots.

   (a) After running the simulation with default parameters, what are the z-statistic and p-value? What do these values indicate about the performance difference between the two groups?

   The z-statistic is 5.97, and the p-value is 1. The z-statistic indicates a significant performance difference between the two groups, with students using cheat sheets performing better on average. The p-value of 1 indicates that the observed difference is highly statistically significant, so the performance difference is not due to random chance.

   (b) How do the results change when you increase the mean score of students using cheat sheets? Explain the impact on the z-statistic and p-value.

   Increasing the mean score of students using cheat sheets will increase the z-statistic and decrease the p-value. This indicates a stronger performance difference between the two groups, making the results more statistically significant.

   (c) How does changing the standard deviation of scores for students using cheat sheets affect the results? What does this imply about the consistency of performance within the group?

   Changing the standard deviation of scores for students using cheat sheets will impact the z-statistic and p-value. A higher standard deviation will result in a lower z-statistic and a higher p-value, indicating less consistency in performance within the group. This implies that students using cheat sheets have varying levels of performance, making it more challenging to detect a significant difference between the two groups.

   (d) Compare the impact of changing the standard deviation for both groups on the results. Which group's variability has a more significant impact on the z-test outcome?

   The standard deviation of the group with the smaller sample size will have a less significant impact on the z-test outcome. This is because the standard error is directly proportional to the pooled standard deviation, meaning that smaller sample sizes are less sensitive to changes in standard deviation.

   (e) How does increasing the total number of students impact the z-test results? Why does the sample size matter in statistical tests?

   Increasing the total number of students will decrease the standard error and increase the z-statistic, making the results more statistically significant. Sample size matters in statistical tests because it affects the precision and reliability of the estimates. Larger sample sizes provide more accurate estimates of the population parameters and reduce the margin of error in the results.

   (f) Discuss the implications of having an unequal number of students in the two groups. How does this affect the reliability of the test results?

   Having an unequal number of students in the two groups can introduce bias and affect the reliability of the test results. The group with the smaller sample size will have a larger standard error, making the results less reliable. It is essential to have equal sample sizes in both groups to ensure that the test results are valid and unbiased.