Random Vectors, Sums of RVs and Estimation with Python Simulations

Applied Stochastic Processes: HW 3

Due Date: October 20, 2024

Objectives

This assignment is designed to help students develop a deep understanding of probability and random variables. Students will enhance their problem-solving skills and gain practical experience through Python-based simulations.

Policy

- You can discuss HW problems with other students, but the work you submit must be written in your own words and not copied from anywhere else. This includes codes.
- However, do write down (at the top of the first page of your HW solutions) the names of all the people with whom you discussed this HW assignment.
- You may decide to write out your solution with pen and paper and use a scanning app to turn in a PDF submission or may choose to type out your solution with LATEX. We strongly encourage the latter.

Expected Topics Covered

1. Random Vectors and Principal Component Analysis
2. Sum of Random Variables and Probability Bounds
3. Estiamtion and Hypothesis Testing

Submission Guidelines

- Submit your code solutions as a Jupyter notebook file (.ipynb) or as Python scripts (.py). You can also convert them to pdf and attach them to your final submission document. Be sure to indicate which code belongs to what question.
- For those using **LaTeX**, you can paste your code by importing the python environment `pythonhiglights`.

```
\usepackage{pythonhiglights}

\begin{python}
# import necessary libraries
import math
import random

# example arithmetic evaluation
a, b = 2, 3
c = a + b
print(f"The sum of the numbers {a} and {b} is {c}")
\end{python}
```

- Ensure all code is well-documented and includes comments explaining each step.
- Provide a brief report summarizing your findings and the results of your simulations.

Q1: Random Vectors and Principal Component Analysis (55 points)

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 $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ where each X_i is a random variable. The covariance matrix of $\mathbf{X}$ 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 $x_1, x_2, x_3 \in \{1, 2, \ldots\}$ and $0 < p < 1$.

   Find the joint PMF $P_Y(y)$ of the transformed random vector Y.

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 platform users. The covariance matrix of this random vector is:

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

Tasks

1. Interpretation of the Covariance Matrix:

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

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

2. Random Vector and Variance:

(a) **(2 points)** What is the total variance of the random vector X?

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

3. Eigenvalues and Eigenvectors of the Covariance Matrix:

(a) **(4 points)** Calculate the eigenvalues and eigenvectors of the covariance matrix Σ_X by hand.

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

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 the random vector.

Tasks

1. 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?

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

2. Transformation of Random Vector:

Let the eigenvector matrix be P and define the transformed random vector Y by $Y = P^T X$.

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

(b) **(2 points)** How does this transformation affect the correlation between the transformed features?

Part 3: Performing PCA by Hand on a Simple Dataset (8 points)

Consider a simple dataset represented by the following 2-dimensional random vector $Y = [Y_1, Y_2]^T$:

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

Tasks

1. Step 1: Mean Centering: (2 points)

(a) Compute the mean of the dataset for each feature Y_1 and Y_2.

(b) Subtract the mean from each feature to center the data.

2. Step 2: Covariance Matrix: (2 points)

(a) Calculate the covariance matrix for the centered dataset.

3. Step 3: Eigenvalue Decomposition: (2 points)

(a) Manually compute the eigenvalues and eigenvectors of the covariance matrix.

(b) Identify the principal components by determining which eigenvalue is larger.

4. Step 4: Project the Data: (2 points)

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

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

Part 4: PCA in Practice with a Larger 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, and Activity Type 2. You will apply PCA using Python to reduce the dimensionality.

Tasks

1. Step 1: Load the Dataset:

(a) Load the dataset using Pandas.

(b) Compute the sample covariance matrix of the data.

2. Step 2: Perform PCA:

(a) Perform PCA on the dataset using Python and calculate the eigenvalues and eigenvectors.

(b) Determine how many principal components are needed to retain at least 90% of the total variance.

3. Step 3: Project the Data:

(a) Project the data onto the first two principal components.

(b) Create a 2D scatter plot of the transformed data.

4. Step 4: Interpretation of Results:

(a) Based on the scatter plot, explain whether the data points are well-separated along the two principal components.

(b) How well does the 2D representation capture the original structure of the data?

Part 5: Interpretation and Business Insights (20 points)

Tasks

1. Feature Interpretation in PCA:

(a) **(5 points)** Based on the principal component directions, explain which features (original dimensions) contribute most to the first and second principal components.

(b) **(5 points)** How would you explain the reduced features to a non-technical team in terms of user behavior patterns?

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?

(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?

Q2: Sum of Random Variables, Central Limit Theorem and Probability Bounds (50 points)

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: $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 + 2\text{Cov}(X, Y)$.
This concept applies to any number of random variables. For independent variables $X_1, X_2, \ldots, X_n$, the sum $Z = X_1 + X_2 + \ldots + X_n$ 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 $\frac{1}{k^2}$. 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 telecommunications 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:

$$P(V) = 0.6 \quad (60\% \text{ voice calls})$$
$$P(D) = 0.4 \quad (40\% \text{ data calls})$$

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 total calls.

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

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

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

(d) **(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.

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

Part 2: Chernoff Bound and 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:

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

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.

(a) **(3 points)** Derive the moment-generating function $\phi_Y(s)$.

(b) **(5 points)** Find $E[Y]$ and $\text{Var}(Y)$, the expected total points and variance.

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

Part 4: Course Enrollment and Resource Planning (6 points)

The number of students enrolling in a popular Data Science course is modeled as a Poisson random variable with a mean of 100 students. The professor has decided that if 120 or more students enroll, he will split the class into two sections, otherwise, he will teach all the students in a single section.

(a) **(3 points)** What is the probability that the professor will need to teach two sections?

(b) **(3 points)** Based on this 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?

Part 5: Comparison of Markov, Chebyshev, and Chernoff Inequalities (19 points)

Consider a Gaussian random variable X with mean $\mu = 0$ and standard deviation $\sigma = 1$. We are interested in comparing how well three probability bounds—**Markov's inequality**, **Chebyshev's inequality**, and the **Chernoff bound**—estimate the probability that X exceeds a given threshold c, i.e., $P(X \geq c)$.

a. Markov's Inequality

For a non-negative random variable X, Markov's inequality provides the following bound:

$$P(X \geq c) \leq \frac{\mathbb{E}[X]}{c}.$$

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

b. Chebyshev's Inequality

Chebyshev's inequality, applicable to any random variable X with variance σ^2, gives a bound for deviations from the mean:

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

We will apply this inequality to the Gaussian random variable.

c. Chernoff Bound

For a Gaussian random variable X, the Chernoff bound provides a tighter bound for tail probabilities:

$$P(X \geq c) \leq \exp\left(-\frac{(c-\mu)^2}{2\sigma^2}\right).$$

This bound is especially useful for normally distributed data.

Tasks

1. **(6 points) Simulation:**
   - Generate $n \in [1000, 100000]$ samples with step = 1000 from a Gaussian distribution $N(0, 1)$.
   - For a given threshold $c \in [0.5, 3.0]$ with step of 0.1, calculate the empirical probability $P(X \geq c)$ based on the generated samples.
   - Compare this empirical probability with the bounds provided by Markov's inequality, Chebyshev's inequality, and the Chernoff bound.
2. **(5 points) Visualization:**
   - Plot a bar chart to compare the empirical probability with the three bounds for different values of c.
   - Allow dynamic visualization to adjust the threshold c and sample size n.
3. **(8 points) Inferential Questions:**
   - Which of the three inequalities provides the tightest bound for different values of c?
   - How does the performance of the bounds change as c increases?
   - What are the advantages and disadvantages of each inequality when applied to this problem?
   - How does the sample size n affect the accuracy of the empirical probability compared to the bounds?

Step-by-Step Guide for the Simulation

1. **Generate Samples:** Use Python's `numpy` library to generate n samples from a Gaussian distribution $N(0, 1)$.
2. **Compute Empirical Probability:** Count how many samples exceed the threshold c and divide by n to estimate $P(X \geq c)$.
3. **Calculate Bounds:**
   - **Markov Bound:** Compute $\frac{\mathbb{E}[X]}{c}$, where $\mathbb{E}[X]$ is the mean of the samples.
   - **Chebyshev Bound:** Use $\frac{\sigma^2}{(c-\mu)^2}$ where σ^2 is the variance.
   - **Chernoff Bound:** Compute $\exp\left(-\frac{(c-\mu)^2}{2\sigma^2}\right)$.
4. **Plot Results:** Plot the empirical probability and the three bounds on the same chart using `matplotlib`.
5. **Dynamic Visualization:** Use `ipywidgets` to allow interactive adjustments of c and n, dynamically updating the plot.

Q3: Estimation and Hypothesis Testing (130 points)

Reading: Estimation and Hypothesis Testing provide a framework for making inferences 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 estimation:

1. **Point Estimation**: Provides a single value as the estimate of a population parameter.
   - Example: The sample mean ($\bar{X}$) estimates the population mean (μ).
   - Formula:
   $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
2. **Interval Estimation**: Offers a range of values, known as a **confidence interval**, believed to contain the parameter with a specified level of confidence (e.g., 95% or 99%).
   - For a population mean, the confidence interval (CI) can be calculated as:
   $$\text{CI} = \bar{X} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$
   - Where $Z_{\alpha/2}$ is the z-score corresponding to the desired confidence level.

Hypothesis Testing is a statistical method used to make probabilistic decisions about population parameters based on sample data. It involves the following steps:

1. **Formulate Hypotheses**:
   - Null Hypothesis (H_0): Represents the default assumption (e.g., no effect or no difference).
   - Alternative Hypothesis (H_1): Represents the claim to be tested.
2. **Select a Significance Level (α)**: Common values are 0.05 or 0.01, which define the threshold for rejecting H_0.
3. **Calculate the Test Statistic**:
   - For testing a mean difference:
   $$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$
   - Where μ_0 is the population mean under the null hypothesis.
4. **Decision Rule**:
   - Compare the test statistic to the critical value from the relevant distribution (e.g., Z-table for normal distribution).
   - If $|Z| > Z_{\alpha/2}$, reject H_0.
5. **Outcome**:
   - 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.

The **Central Limit Theorem (CLT)** plays a crucial role in both estimation and hypothesis testing, stating that the sampling distribution of 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 interpretations in estimation and hypothesis testing.

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

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:

$$8.2, 8.5, 8.9, 9.0, 7.8, 8.6, 8.4, 8.1, 9.1, 8.7, 9.2, 8.8, 8.3, 9.3, 8.0, 8.9, 8.4, 8.6, 8.7, 8.2$$

1. **(2 points)** Calculate the **sample mean** and **sample variance** .
2. **(3 points)** What does the sample mean estimate in this case, and determine if the sample mean is an unbiased estimator of the population mean μ. Explain your reasoning.
3. **(3 points)**Compute the mean squared error (MSE) of the sample mean, which is defined as

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

4. **(2 points)**Explain how the **error of the mean** and the sample variance might change if the sample size were smaller or larger.
5. **(2 points)** What effect do outliers have on the sample mean and variance?
6. **(8 points)** Dynamically simulate different sample sizes $n \in [5, 1000, \texttt{step=5}]$ 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 points)

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 the **95% confidence interval** for the population mean height, assuming that the population standard deviation is unknown.
2. **(2 points)** If the population standard deviation was known to be 7 cm, how would this change the confidence interval?
3. **(2 points)** What does the confidence interval mean in practical terms?
4. **(2 points)** How does increasing the sample size to 100 men affect the confidence interval?
5. **(8 points)** Simulate the construction of confidence intervals for different sample sizes and confidence levels. Vary $nin[10, 500, \texttt{step=10}]$ and $\alpha in[0.01, 0.2\texttt{step=0.01}]$ dynamically and observe how the confidence interval width changes.

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.
2. **(6 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.
3. **(2 points)** What does the **p-value** mean in the context of this A/B test?
4. **(3 points)** How would you interpret the results to your marketing team?
5. **(6 points)** Use a dynamic simulation to vary the sample sizes $n \in [100, 1000, \texttt{step=50}]$ and click-through $c \in [50, 600, \texttt{step=10}]$ rates for Versions A and B. Observe how the **p-value** and hypothesis test decision change as you modify these parameters.

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.
2. **(4 points)** Compute the **sensitivity**, **specificity**, **positive predictive value (PPV)**, and **negative predictive value (NPV)** of the test.
3. **(4 points)** What do these values indicate about the **reliability** of the medical test?
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?
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.

Simulation Steps

Step 1: Adjust Test Parameters

For the simulation, we define the following ranges and step sizes for each parameter:

- TP: Range [0, 500], Step size: 10
- FP: Range [0, 500], Step size: 10
- TN: Range [0, 500], Step size: 10
- FN: Range [0, 500], Step size: 10

Step 2: Compute Metrics

Based on the values of TP, FP, TN, and FN, compute the performance metrics are computed:

Step 3: Visualize Using Bar Plots

To visualize the results of the simulation, generate a bar plot is showing the values of:

- Sensitivity
- Specificity
- PPV
- NPV

Plot the metrics dynamically as the values of TP, FP, TN, and FN are adjusted. The bar plot updates in real-time to reflect the changes in the test's performance.

Part 5: MLE / MAP Estimation (14 points)

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):

H1: 0%	H2: 15%	H3: 30%	H4: 45%
H5: 60%	H6: 75%	H7: 90%	H8:100%

Table 1: Coin toss observed under various hypothesis

1. **(4 points)** Given the first five coin flips: $[1, 0, 1, 1, 1]$, calculate the likelihood of these observation under each hypothesis.

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.

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

   (c) What is the most likely hypothesis after all observations are made?

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)

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%

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?

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

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?

   (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.

   (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?

   (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?

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

   (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?

Steps for Simulation and Analysis

We aim to simulate an examination process where two groups of students are compared: one group uses cheat sheets, and the other does not. We will perform a z-test and adjust parameters interactively to observe their impact.

Step 1: Define Parameters

The parameters to simulate the exam process are:

- **Mean score of students using cheat sheets** (μ_1): range [85, 95], step size: 1
- **Mean score of students not using cheat sheets** (μ_2): range [80, 90], step size: 1
- **Standard deviation of students using cheat sheets** (σ_1): range [1, 5], step size: 0.1
- **Standard deviation of students not using cheat sheets** (σ_2): range [1, 5], step size: 0.1
- **Sample size for both groups** (n_1, n_2): range [20, 200], step size: 10

Step 2: Calculate Pooled Standard Deviation and Standard Error

Using the provided formulas, calculate the pooled standard deviation (S_p) and standard error (SE) for each set of parameter values.

Step 3: Perform Z-test

For each set of simulated parameters, calculate the z-score as:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{SE}$$

Compare the z-score to the critical value for a one-tailed test ($z_{crit} = 1.645$ for $\alpha = 0.05$).

Step 4: Plot Results Using Bar Plots

Plot the z-scores and p-values as the parameters vary. Use interactive sliders to dynamically adjust the mean, standard deviation, and sample size.