Question 1

In each part below, we consider the result of $n$ independent tosses of a fair coin, for which $P(\text{Heads}) = P(\text{Tails}) = 0.5$. Numerically evaluate and report all bounds and probabilities.

(a) Let $X$ be the number of heads observed in $n = 10$ independent coin tosses. Use Chebyshev’s inequality to lower bound $P(4 \leq X \leq 6)$.

(b) Let $X$ be the number of heads observed in $n = 100$ independent coin tosses. Use Chebyshev’s inequality to lower bound $P(40 \leq X \leq 60)$.

(c) Let $X$ be the number of heads observed in $n = 1000$ independent coin tosses. Use Chebyshev’s inequality to lower bound $P(400 \leq X \leq 600)$.

(d) Using the binomial cumulative distribution function, evaluate the exact probabilities of the events in parts (a,b,c). Use the Matlab \texttt{binocdf} function to do this in a numerically stable way. Compare the Chebyshev bounds to the true probabilities, and discuss any trends.

(e) Using a normal approximation, obtain an asymptotic approximation of the events in parts (a,b,c). Concretely, replace the binomial distribution with a Gaussian distribution with mean and variance exactly equal to the mean and variance of the binomial distribution, and compute Gaussian tail bounds with the Matlab \texttt{normcdf} function. Compare these bounds to the Chebyshev and binomial tail bounds, and discuss any trends.

Answer 1
Question 2

Let $X$ be a non-negative random variable with expected value $E[X] = 4$, and let $a > 0$ be some positive number. We derive and evaluate various upper bounds on $P(X \geq a)$.

(a) Suppose that $0 < a < E[X]$. What is the best upper bound you can give for $P(X \geq a)$?

(b) Use Markov’s inequality to give an upper bound on $P(X \geq a)$ that is valid for any $a > 0$.

(c) For any non-negative random variable, the proof of Markov’s inequality (from lecture and the textbook) can be slightly generalized to show that

$$P(X \geq a) \leq \frac{E[X^2]}{f(a)},$$

where $f(a) > 0$ is some constant that depends on $a$. Prove this bound and find $f(a)$.

(d) Suppose that in addition to knowing $E[X] = 4$, you know that $Var(X) = 10$. Apply the inequality from part (c) to upper bound $P(X \geq a)$. For which values of $a > 0$ is this bound better (tighter) than the Markov’s inequality bound from part (b)لة؟

Answer 2
Question 3

Eli and Lorenzo are playing a game. They first each use a program that independently generates a continuous random number uniformly distributed between 0 and 1. Let \( X \) be Eli’s number, and \( Y \) be Lorenzo’s number. Recall that the correlation coefficient between any two random variables \( A \) and \( B \) equals

\[
\rho(A, B) = \frac{\text{Cov}(A, B)}{\sqrt{\text{Var}(A)\text{Var}(B)}}.
\]

Below we use the correlation coefficient to determine the strength of dependence between the raw scores of Eli and Lorenzo, and various summaries of these scores.

(a) Determine \( \rho(X, Y) \), the correlation between Eli’s score \( X \) and Lorenzo’s score \( Y \).

(b) Let \( Z = X + Y \), the sum of their individual scores. Determine \( \rho(X, Z) \), the correlation between Eli’s score \( X \) and the overall score \( Z \).

(c) Let \( Z = X + Y \) and \( W = X - Y \), the sum and difference of their individual scores. Determine \( \rho(Z, W) \), the correlation between this sum and difference.

(d) Consider a game where the higher-scoring player wins points equal to the sum of the scores of the two players. Thus Eli’s winnings \( Z = 0 \) if \( X < Y \), and \( Z = X + Y \) if \( X \geq Y \). Determine \( \rho(X, Z) \).

Answer 3
Question 4

In this problem, your goal is to predict the horsepower $Y$ of a car based on observation of some other feature $X$ of that car. The input feature $X$ and output response $Y$ are both real numbers, so we will use models based on bivariate normal distributions. The training data, and corresponding bivariate normal approximations, are plotted in Figure 1.

Suppose that $E[X] = \mu_x$, $E[Y] = \mu_y$, $Var(X) = \sigma_x^2$, $Var(Y) = \sigma_y^2$, and $\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$ is the correlation coefficient. We have provided code that estimates these means, variances, and covariances based on the empirical distribution of the training data. For some test data where $X = x$, you will then predict $Y$ via its conditional mean:

$$\hat{y} = E[Y|X = x] = \mu_y + \frac{\rho \sigma_y}{\sigma_x} (x - \mu_x)$$

(1)

Given a test dataset of $M$ observations $(x_i, y_i)$, and a model that predicts $\hat{y}_i = E[Y|X = x_i]$ for true response $y_i$, we measure the root mean squared error in our predictions as follows:

$$L(y, \hat{y}) = \sqrt{\frac{1}{M} \sum_{i=1}^{M} (y_i - \hat{y}_i)^2}$$

(2)

Our goal is to create models that make this error as small as possible. For full credit, the specified error values and plots must be included in your solution pdf.

Figure 1: Scatter plots of $Y = \text{horsepower}$ versus $X = \text{weight}$ (left), and $Y = \text{horsepower}$ versus $X = \text{miles per gallon (MPG, right)}$, for the automobile training data. For each dataset, we plot contours of constant probability for a bivariate normal distribution fit to the data.

(a) We have provided a function `pred_linear` that predicts horsepower $Y$ as the mean of the training data, $\hat{y} = \mu_y$, ignoring the input features $x$. Improve this function so that it predicts $Y$ using the conditional mean of Equation (1).

(b) Let the input feature $X$ be vehicle weight, and apply your `pred_linear` method to predict the test vehicles horsepower. What is the root mean squared error of your predictions? Plot the true and predicted horsepower values for each test example.
(c) Let the input feature $X$ be vehicle MPG, and apply your \texttt{pred\_linear} method to predict the test vehicles horsepower. What is the root mean squared error of your predictions? Plot the true and predicted horsepower values for each test example.

From visual inspection, the distribution of the MPG data $X$ does not seem to fit our Gaussian assumption. Fixing a threshold of $t = 20$ MPG, we will explore whether we can build a more accurate model by splitting the data into two groups. The first group contains all the examples where $X \leq t$, and the second the examples where $X > t$. Using the training data, we estimate separate bivariate normal distributions for each of the two groups. For a test input $x_i$, we check whether $x_i \leq t$ or $x_i > t$, and predict $\hat{y}_i$ using the corresponding normal.

(d) Implement the threshold-based regression method described above. Let the input feature $X$ be vehicle MPG, the threshold $t = 20$ MPG, and apply your method to predict the test vehicles horsepower. What is the mean squared error of your predictions? Plot the true and predicted horsepower values for each test example. Compare the accuracy of this approach to the regression methods in parts (b) and (c).

\textbf{Answer 4}