Question 1

Let $X, Y$ be uniform random variables chosen independently from the interval $[0,1]$. Find the cumulative distribution function and density for the following r.v.'s:

(a) $A = |X - 1/2|
(b) B = (X - 1/2)^2
(c) C = |X - Y|
(d) D = \max(X, Y)
(e) Find the expected value of $D$.

Answer 1
Question 2

A random variable $X$ has a Poisson distribution of parameter $\lambda$ ($X \sim \text{Poisson}(\lambda)$) if the distribution is given by

$$P(X = n) = \frac{e^{-\lambda} \lambda^n}{n!}, m = 0, 1, 2...$$

(a) Use the Taylor series for $e^x$ to find the moment generating function of the r.v. ($X \sim \text{Poisson}(\lambda)$). Prove that $\lambda$ is the expectation of the distribution.

(b) Show that if ($X \sim \text{Poisson}(\lambda)$) and $X_1, ..., X_n$ are independent and identically distributed with ($X_1 \sim \text{Poisson}(\frac{\lambda}{n})$) then $X$ and $S_n = \sum_{x=1}^{n} X_i$ have the same distribution.

(c) Use the previous item to find a formula for $P(S_n \leq n)$.

Answer 2
Question 3

Eli and Cyrus 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 Cyrus’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 Cyrus, and various summaries of these scores.

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

Answer 3
Question 4

In this problem, your goal is to predict the calories $Y$ of a MacDonald’s meal based on observation of some other feature $X$ of that meal. 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.

(a) We have provided a function `pred.linear` that predicts calorie value $Y$ 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 the fat content of the meal, and apply your `pred.linear` method to predict the test meal’s calorie content. What is the root mean squared error of your predictions? Plot the true and predicted calorie values for each test example.

(c) Let the input feature $X$ be carbohydrate content of the meal, and apply your `pred.linear` method to predict the test meal’s calorie content. What is the root mean squared error of your predictions? Plot the true and predicted calorie values for each test example.

All data for this problem comes from [https://www.kaggle.com/mcdonalds/nutrition-facts](https://www.kaggle.com/mcdonalds/nutrition-facts).
Figure 1: Scatter plots of $Y = \text{Calorie content}$ versus $X = \text{Fat content}$ (left), and $Y = \text{Calorie content}$ versus $X = \text{Carbohydrate content}$ (right), for the MacDonald’s training data. For each dataset, we plot contours of constant probability for a bivariate normal distribution fit to the data.