**Objective:** Suppose that $Y = Y(X_1, X_2)$ is a random variable whose distribution depends on two independent variables $X_1$ and $X_2$, and for any given pair $(X_1, X_2)$, it is an important to investigate the mean $\mu = \mathbf{E}(Y)$ and variance $V = Var(Y)$. The objective of this midterm exam is to estimate the mean $\mathbf{E}(Y) = \mu(X_1, X_2)$ and the variance $V = Var(Y) = V(X_1, X_2)$ as deterministic functions of $X_1$ and $X_2$ when $0 \le X_1 \le 7$ and $0 \le X_2 \le 4$.

**Background:** The motivation of the random variable $Y = Y(X_1, X_2)$ is from overshoot analysis of random walks in applied probability. For a pair of given $X_1$ and $X_2$, one first generates a sequence of stationary (dependent) positive-valued random variables, $Z_1, Z_2, \cdots, Z_m$, say, $m = 50$, whose joint distribution depends on both $X_1$ and $X_2$. Then the random variable $Y = Y(X_1, X_2)$ is defined as

$$Y = Y(X_1, X_2) = \frac{\max(Z_1, Z_2, \cdots, Z_m)}{Z_1 + Z_2 + \cdots + Z_m}. \tag{1}$$

Clearly $0 \le Y \le 1$, but it is an open problem to characterize the exact distribution of $Y = Y(X_1, X_2)$ in general as a function of $X_1$ and $X_2$.

For your information, no research has been done to investigate on the properties of $Y = Y(X_1, X_2)$ in (1), and by asymptotic analysis when $m$ goes to $\infty$, one conjecture on the mean of $Y$ is that

$$\mu(X_1, X_2) = \mathbf{E}(Y(X_1, X_2)) \approx g(\frac{1 + X_2}{2\sqrt{1 + X_1}}), \qquad \text{where} \quad g(w) = \exp\Big( -2 \sum_{n=1}^{\infty} \frac{1}{n} \Phi(-\frac{\sqrt{n}}{2} w) \Big). \tag{2}$$

Here $\Phi(x) = P(N(0,1) \le x)$ is the cdf of the standard normal random variable, and $g(w)$ is a real-valued function that appears quite frequently in renewal theory in applied probability. Please note that this conjecture may or may not be correct, and you may or may not explore this conjecture in your analysis below.

**Training data set:** In order to help you to develop a reasonable estimation of the mean and variance of $Y = Y(X_1, X_2)$ as deterministic functions of $X_1$ and $X_2$, we provide a training data set that is generated as follows. We first choose the uniform design points when $0 \le X_1 \le 7$ and $0 \le X_2 \le 4$, that is, $x_{1i} = 0.1 * i$ for $i = 0, 1, \ldots, 70$, and $x_{2j} = 0.1 * j$ for $j = 0, 1, \ldots, 40$. Thus there are a total of $(70 + 1) * (40 + 1) = 2911$ combinations of $(x_{1i}, x_{2j})$'s, and for each of these $71 * 41 = 2911$ combinations, we generate 200 independent sequences of $(Z_1, Z_2, \cdots, Z_m)$, thereby yielding 200 independent realizations of $Y_{ijk}$ from Equation (1) for $k = 1, \ldots, 200$.

The corresponding training data, `midtermtrain.csv`, is available from T-square or the following R code

```
midtermtrain <- read.table(file = "http://www2.isye.gatech.edu/~ymei/7406/midtermtrain.csv", sep=",");
dim(midtermtrain);
```

Note that the dimension of the training data set "`midtermtrain`" is $2911 \times 202$. Each row corresponds to one of $71 * 41 = 2911$ combinations of $(X_1, X_2)$'s. The first and second columns are the $X_1$ and $X_2$ values, resp., whereas the remaining 200 columns are the corresponding 200 independent realizations of $Y$'s.

Based on the training data, you are asked to develop an accurate (and ideally still simple) estimation of the mean and variance function of $Y = Y(X_1, X_2)$, i.e., the functions $\mu(X_1, X_2)$ and $V(X_1, X_2)$, as deterministic functions of $X_1$ and $X_2$ when $0 \le X_1 \le 7$ and $0 \le X_2 \le 4$.

**Testing data set:** For the purpose of evaluating your proposed estimation models and methods, we choose 26 design points for $X_1 = 0.10536, 0.22314, \ldots, 5.92370, 6.61690$, and uniform design points for $X_2 = 0.1 * j$ with $j = 0, 1, \ldots, 40$. Thus there are a total of $26 * 41 = 1066$ combinations of $(X_1, X_2)$ in the testing data set. You are asked to use your formula to predict the mean and variance of $Y = Y(X_1, X_2)$ for the $26 * 41 = 1066$ combination of $(X_1, X_2)$ in the testing data (please keep the six digits for your answers).

The exact values of the $(X_1, X_2)$'s in the testing data set are included in the file `midtermtrain.csv`, which is available from T-square or the following R code

```
midtermtest  <- read.table(file = "http://www2.isye.gatech.edu/~ymei/7406/midtermtest.csv", sep=",");
```

**Estimation Evaluation Criterion:** In order to evaluate your estimation or prediction, I also generated 200 random realizations of $Y$'s for each combination of $(X_1, X_2)$ in the testing data set. However, I will not release these 200 independent realizations for the testing data. Instead I will use them to compute a baseline estimation of $\hat{\mu}_{base}(X_1, X_2)$ and $\hat{V}_{base}(X_1, X_2)$ for the testing data. Specifically, for each given combination of $(X_1, X_2)$, we have 200 realizations of $Y$'s, denoted by $Y_1, \cdots, Y_{200}$, and then we compute the baseline estimations as the Monte Carlo "sample" mean and variance:

$$\hat{\mu}_{base} = \bar{Y} = \frac{Y_1 + \cdots + Y_{200}}{200} \quad \text{and} \quad \hat{V}_{base} = \frac{1}{200-1}\sum_{i=1}^{200}(Y_i - \bar{Y})^2.$$

Your predicted mean or variance functions, say, $\mu^*(X_1, X_2)$ and $V^*(X_1, X_2)$, will then be evaluated as compared to the baseline Monte Carlo estimations, $\hat{\mu}_{base}(X_1, X_2)$ and $\hat{V}_{base}(X_1, X_2)$:

$$MSE_\mu = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}(\mu^*(x_{1i}, x_{2j}) - \hat{\mu}_{base}(x_{1i}, x_{2j}))^2$$
$$MSE_V = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}(V^*(x_{1i}, x_{2j}) - \hat{V}_{base}(x_{1i}, x_{2j}))^2, \tag{3}$$

where $(I, J) = (26, 41)$ for the testing data.