Question 1 (suggested for peer/personal tutorial)

Consider the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Recall* the definition of an indicator variable for an event $A \in \mathcal{F}$, denoted $\mathbb{I}_A$ (or $\mathbb{I}(A)$) and defined for $\omega \in \Omega$ by

$$\mathbb{I}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A. \end{cases}$$

(*See Definition 7.3.2 from Prof. Veraart's notes in Term 1.)

(a) Is $\mathbb{I}_A$ a discrete random variable or a continuous random variable?

(b) If $\mathbb{I}_A$ is discrete, write down its probability mass function, or if it is continuous, write down its probability density function.

(c) Compute $\mathrm{E}(\mathbb{I}_A)$.

Question 2

Prove Theorem 1.1.2 from the notes:

> **Theorem 1.1.2.** Given two arbitrary random variables X and Y with a specified joint distribution, suppose that X and Y both have finite means. Then the function g of X that minimises $\mathrm{E}[(Y - g(X))^2]$ is $g(X) = \mathrm{E}[Y|X]$, i.e.
> $$\min_g \mathrm{E}[(Y - g(X))^2] = \mathrm{E}[(Y - \mathrm{E}[Y|X])^2]$$

Question 3

Given a random variable X, the median of the distribution of X is a value m such that $\mathrm{P}(X \le m) \ge \frac{1}{2}$ and $\mathrm{P}(X \ge m) \ge \frac{1}{2}$. Show that if X is a **continuous** random variable with probability density function $f_X(x)$ then

$$\min_c \mathrm{E}(|X - c|) = \mathrm{E}(|X - m|).$$

Question 4 (Challenge)

Question 3 proves the result in a special case, since it is assumed that X is continuous and it is assumed that a p.d.f. exists. Prove the result in general:

$$\min_c \mathrm{E}(|X - c|) = \mathrm{E}(|X - m|),$$

when X is either discrete or continuous, only using the properties of the expectation. As before, m is the median of the distribution of X.

Hints:

- Question 2: It might be useful to use the "alternative" notation, e.g. $\mathrm{E}_X[X]$, and use the result in Exercise 1.1.6 of the notes: $\mathrm{E}_X[g(X)h(Y)] = g(X)\mathrm{E}_X[h(Y)]$.
- Question 3: Reformulate the definition of a median using f_X and recall the Leibniz integral rule:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\int_{a(t)}^{b(t)} g(x,t)\mathrm{d}x\right) = g\left(b(t),t\right)\cdot\left(\frac{\mathrm{d}}{\mathrm{d}t}b(t)\right) - g\left(a(t),t\right)\cdot\left(\frac{\mathrm{d}}{\mathrm{d}t}a(t)\right) + \int_{a(t)}^{b(t)}\left(\frac{\partial}{\partial t}g(x,t)\right)\mathrm{d}x.$$

For a nice reference for the Leibniz integral rule, see the first few pages of H. Flanders. Differentiation under the integral sign. *The American Mathematical Monthly*, 80(6):615-627, 1973.

Question 5 (using R)

Suppose there is a file named `file1.txt` which contains the following data:

```
x,y
2,3
4,6
6,9
8,12
```

(Either download the file from Blackboard, or copy-paste the data into a file and name it `file1.txt`.)

(a) Use the function `read.table` to read the data from `file1.txt` into a data frame object named `df`.

(b) Extract a vector named `x`, containing values $(2, 4, 6, 8)$ from the data frame `df`. Similarly, extract a vector named `y`, containing values $(3, 6, 9, 12)$ from the data frame `df`.

(c) Create a vector named `z` which is the mean of the two vectors `x` and `y`, i.e. `z` contains four values, the first of which is $(2 + 3)/2 = 2.5$.

(d) Add the vector `z` to the data frame `df` so that `df` contains three columns, `x`, `y` and `z`.

(e) Write the data frame `df` to a file named `file2.txt`, so that this file contains:

```
x,y,z
2,3,2.5
4,6,5
6,9,7.5
8,12,10
```

Hints:

- Question 5(a): You will need to set the `sep` parameter of the function `read.table` to have value `","`, and you will also need to set the `header` parameter of the function `read.table` to be `TRUE`.
- Question 5(e): Use the function `write.csv` and set the `quote` and `row.names` parameters of the function to be `FALSE` (or `F`).