03. Estimators

To be completed Friday 18th of November, by 12 noon sharp. You are encouraged to submit in pairs, but are also allowed to submit alone. **One person** must submit the solution as **a single pdf file**, to the folder on Ilias, by naming *exercise_03_name1_matriculationnumber1_name2_matriculationnumber2.pdf*, with the obvious replacements in the strings. **Not adhering to the formatting requirements may result in the submission not being graded.**
Exercises are only graded in a binary fashion as sufficient or insufficient. To be graded as sufficient, you do not necessarily have to have correct solutions to every sub-question, but you must have made a clear and earnest effort to solve the entire exercise. Ultimately, what constitutes sufficient is at the discretion of the tutors. To be admitted to the exam, you must have submitted sufficient answers to at least 5 of the 6 (maybe 7) exercise sheets.

1. **EXAMple Question — The importance of Importance Sampling**
   Let's say we want to understand the relationship between attending lectures and passing the respective exams throughout the University of Tübingen. We assume that attending and passing are binary events (a student either attends all the lectures of a course or skips them altogether, and either passes—with sufficient—or fails the exam—not sufficient—, independent of the other students). In order to understand this relationship, we ran a survey in the Data Literacy course after grading the exam by asking the students whether they attended the lectures. 100 students responded in the following way:

   Table 1: Survey results.

	Attend (A)	Skip ($\neg A$)
Sufficient (S)	56	5
Not sufficient ($\neg S$)	10	29

   (a) Using the Sampling Estimator and Table 1, estimate the probabilities $P(S|\neg A)$ and $P(\neg S|A)$.

   What is the variance of the estimators? Note that, as we have seen in the lecture, in order to estimate the variance of the estimators, we need to make the rather unrealistic assumption that we know the expected $P(S|\neg A)$ and $P(\neg S|A)$, so we will assume here that we know that $P(S|\neg A) = 0.1$ and $P(\neg S|A) = 0.2$.

   (b) Overall, at the University of Tübingen, 50% of students attend lectures. Using the data in Table 1 and importance sampling, estimate the probability that students throughout the university pass a course $P(S)$.

   (c) What are some of the challenges of using Importance Sampling?

2. **Theory Question — Unbiased variance estimate**
   In the lecture, we showed that the sampling-based estimators of expectations are unbiased, i.e., that, given i.i.d. data $D = \{x_1, ..., x_N\}$ from the distribution $p(x)$, $\hat{\phi} = \frac{1}{N}\sum_i^N f(X_i)$ is an unbiased estimator for $\phi = \mathbb{E}_p[f(X)]$.

   As a corollary, the sample-average $\hat{\mu} = \frac{1}{N}\sum_i^N x_i$ is an unbiased estimator for the mean, and the sample-variance $\hat{\sigma}^2 = \frac{1}{N}\sum_i^N (x_i - \mu)^2$ is an unbiased estimator for the variance.

   However, it turns out that the 'plug-in' estimate for the variance, $\tilde{\sigma}^2 = \frac{1}{N}\sum_i^N (x_i - \hat{\mu})^2$ *is* biased. In this exercise, we will compute this bias, and how it is often corrected for.

   (a) Show that $\mathbb{E}[\hat{\mu}x_i] = \frac{1}{N}\sum_{j=1}^N \mathbb{E}[x_j x_i]$.

(b) Show that $\mathbb{E}[\hat{\mu}^2] = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[x_i x_j]$.
Hint: $(\frac{1}{N}\sum_{i=1}^{N} x_i)^2 = \frac{1}{N^2}(\sum_{i=1}^{N} x_i)(\sum_{j=1}^{n} x_j)$.

(c) Using 2a and 2b, show that $\tilde{\sigma}^2$ is biased, where $\tilde{\sigma}^2 = \sigma^2\frac{N-1}{N}$, and derive the unbiased estimate.
Hint: $\mathrm{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

3. **Practical Question — Bacterium random-walk**
In this exercise, you will use the *Sampling (Monte-Carlo) estimator* to compute the position of the *E. coli* bacterium, while it performs a *random walk*. See **Exercise_03.ipynb**.