Please carefully read and follow the general instructions regarding exercises. Failing to meet the requirements might lead to penalties. https://moodle.uef.fi/mod/page/view.php?id=1935632

If you suspect that something is wrong with some exercise question, please contact the lecturer.

If you face persistent issues while working on an exercise, do ask for help, e.g. during a course meeting or by contacting the lecturer via email.

**Problem 1** (Variance of mean). Show that if $X_1, \ldots, X_n$ are $n$ independent random variables, each with variance $\sigma^2$, then the variance of their mean $\bar{X}$ equals $\frac{\sigma^2}{n}$ (Note: you should go back to the definition of variance).

**Problem 2** (Sampling for search analytics). A search engine receives a stream of queries and would like to study the behavior of typical users. Assume the stream consists of tuples (user, query, time), and we wish to store at most one tenth of the stream elements. We are interested to know what fraction of the typical user’s queries are repeated over a month (i.e. number of distinct queries issued at least twice by the user divided by the total number of distinct queries issued by the user).

One obvious approach would be to sample one tenth of the incoming queries over the month, and calculate the average number of duplicate queries for a user in the sample. However, this approach does not provide a correct answer.

a) Explain why that is so, considering the case of a user who issues $s$ search queries once during the month, $d$ search queries twice and no queries more than twice. That is, indicate what the correct answer would be for that user, and what number of queries issued once and issued twice we can expect to see repeated and not repeated in the sample if we have a uniform, one-tenth, sample of the queries.

b) Suggest an alternative sampling approach to obtain a correct answer with a one-tenth sample.

**Problem 3** (Bloom filters). We want to build a Bloom filter to store a list of one million email addresses (assume one email address is 20 bytes). We use 5 hash functions and want the probability of false positives to be at most 0.1%.

a) How much storage do we save by using such a Bloom filter as compared to storing the addresses directly?

**Problem 4** (The Gambler). A gambler plays 120 hands of draw poker, 60 hands of blackjack and 20 hands of stud poker per day. He wins a hand of draw poker with probability $1/6$, a hand of black jack with probability $1/2$, and a hand of stud poker with probability $1/5$. Assume the outcomes of the card games are pairwise independent.

a) What is the expected number of hands the gambler wins a day?

b) What would be the Markov bound on the probability that the gambler will win at least 108 hands on a given day?

c) What would the Chebyshev bound be on the probability that the gambler will win at least 108 hands? You may answer with a numerical expression that is not completely evaluated.