DATA11007 Statistics for Data Science
Exam 25 January 2023
Antti Honkela, Mikko Heikkilä and Marlon Tobaben

Answer all problems. Each problem is worth 6 points.

Allowed equipment: writing instruments.

Voit vastata kysymyksiin myös suomeksi.

1. True or false? **Answers to the attached answer sheet.** (Correct answer $\frac{1}{2}$ points, wrong answer $-\frac{1}{4}$ points, empty 0 points)

   i. The variance of the sum of any two random variables is the sum of their variances.
   ii. According to the central limit theorem, the sum $\sum_{i=1}^{n} X_i$ of independent random variables $X_1, X_2, \ldots$, each following the uniform distribution $X_i \sim \text{Uniform}(-1, 1)$, converges in distribution to a Normal distribution.
   iii. The mean squared error (MSE) of an estimator $\hat{\theta}_n$ can be written as

   $$\text{MSE} = \text{bias}(\hat{\theta}_n) + \mathbb{V}(\hat{\theta}_n).$$

   iv. The expected value of a consistent estimator is equal to the true parameter value.
   v. A maximum likelihood estimator is always unbiased.
   vi. Bootstrap sampling can be used to estimate the variance of plug-in estimators.
   vii. p-value is equal to the probability of observing under the null hypothesis a value of the test statistic that is at least as extreme as what was observed.
   viii. Statistically significant differences are always practically important.
   ix. According to the Bayesian method, predictions are computed as averages over the prior distribution.
   x. A minimax estimator always has the smallest Bayes risk for all prior distributions.
   xi. When treatments are randomised, any consistent estimator of association is a consistent estimator of the causal effect.
   xii. Maximising the number of obtained responses is the most important consideration in survey design.

2. Briefly explain the following terms

   i. Confidence interval

   ii. Multiple testing problem

   iii. Simpson's paradox

   iv. Missing not at random (MNAR)