Abteilung Numerik
Jana de Wiljes

Potsdam

2. Problem sheet for **Statistical Data Analysis**

Exercise 1 (3 Points) Let $(x_1, \ldots . x_n) \in \mathbb{R}^n$ be a set of samples. Show that for all $a \in \mathbb{R}$

$$\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 + n(\bar{x}_n - a)^2. \tag{1}$$

Exercise 2 (3 Points) Determine whether S_n is an unbiased estimator of σ. In case it is not an unbiased estimator, which one is larger $\mathbb{E}[S_n]$ or σ?

Exercise 3 (4 Points) Let X_1, X_2, X_3, X_4 be a sample from $U(0, 1)$, and let $X_{(1)}$, $X_{(2)}$, $X_{(3)}$, $X_{(4)}$ be the order statistic. Determine the density of $X_{(1)}$, $X_{(2)}$, $X_{(3)}$, $X_{(4)}$.

Exercise 4 (10 Points) Suppose that there are 10000 students at university of Potsdam and we want to estimate the distribution of the students' height in cm.

1. Generate 10000 random samples from $N(170, 16)$ and assume that the whole set is the population.
2. Take 50 random samples from the population and estimate the mean μ, variance σ^2, standard deviation σ of the distribution of the population. Are they unbiased estimators?
3. Determine Med_n and 31th order statistic.
4. Determine the truncated mean for some k. Is there a k so that the Mean is equal to the median?

Exercise 5 (12 Points)
Suppose that you are running an advertising agency and interested in strategic result-oriented marketing by placing one of the 10 different advertisements at once to one user. The reward is 1 if the ad is clicked by the user, and 0 if not. Assume that the dataset 'Ads.csv' provided on Moodle is the underlying rewards.

1. Visualize the rewards distribution, i.e. the number of being clicked by a user across all the advertisements. What do you observe?
2. Implement Thompson sampling algorithm to maximize the cumulative rewards. Compute the total rewards you could achieve.
3. Implement a function that computes the expected regret using the formula discussed in Lecture 4. Compute the regret.