Probabilistic inference for data science 2 Examination 21.5.2021

1. Consider simple linear regression with predictor x and response y. The data can be found in the file "xy_reg.csv". In R the model can be fitted and estimates extracted using

```
> # Read data
> xy_reg <- read.table("xy_reg.csv", sep=";", header=T)
> # Fit the model on observed data xy_reg
> m <- lm(y~x, data=xy_reg)
> # Extract estimator of the beta-coefficients
> coef(m)
```

   You find variance-covariance matrix of the estimators of β-coefficients with command

```
> vcov(m)
```

   and based on this you can get the correlation matrix of estimator of β's with commands

```
> s.dev <- as.matrix(sqrt(1/diag(vcov(m))))
> vcov(m)*s.dev %*% t(s.dev)
```

   This is how you get the correlation of estimators $\hat{\beta}_0$ (intercept) and $\hat{\beta}_1$ (slope) based on the model m fitted on the data.

   Using bootstrap and 5000 resamples evaluate the correlation of $\hat{\beta}_0$ and $\hat{\beta}_1$. Result should of course be similar to the correlation evaluated based on the model m.

2. Let $X_i = X(u_i)$ be random variable X at spatial locations u_1, u_2 and u_3. Let X_i be normally distributed with $\mathbb{E}(X_i) = \mu$ and $Var(X_i) = 8$ for $i = 1, 2, 3$ (i.e. μ is unknown). The corration between X_i, X_j is defined as $Cor(X_i, X_j) = exp^{-||u_i - u_j||/2}$, where $||u_i - u_j||$ is distance between locations u_i and u_j. Consider locations $u_1 = (1, 0)$, $u_2 = (0, 0)$ and $u_3 = (0, 3)$.

   (a) Present the distribution of

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

   (b) Present the likelihood in the case that we have observed

$$X = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

(c) Find the maximum likelihood estimate of μ when assuming that

$$X = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

has been observed. A numerical approximation is sufficient.

3. Let X be i.i.d. random sample of size $n = 50$ from distribution with expectation $\mu = 10$ and variance $\sigma^2 = 20$. Consider the probability $\mathbb{P}(-1.5 \leq \bar{X}_n - \mu \leq 1.5)$ i.e. the probability that the sample mean $\bar{X}_n$ will not differ from the expectation μ more than 1.5 units.

   (a) Use central limit theorem to approximate probability $\mathbb{P}(-1.5 \leq \bar{X}_n - \mu \leq 1.5)$.

   (b) Use Chebyshev's inequality to approximate the lower bound for the same probability and compare the results.