Exam GEO4300 23.11.20

1 Random parameter estimation

$$X = \begin{cases} -1, & \text{prob.} \ 1/3 \\ 3, & \text{prob.} \ 1/2 \\ 4, & \text{prob.} \ 1/6 \end{cases}$$

a) $E[X] = \sum_{i=1}^{n} x_i f(x_i)$, w/ f as pmf of X.

$$= \sum_{i=1}^{3} x_i f(x_i) = (-1)\cdot\frac{1}{3} + 3\cdot\frac{1}{2} + 4\cdot\frac{1}{6}$$

$$= \underline{\frac{11}{6} \approx 1,83}$$

b) $Var[X] = E[X - E(X)]^2 = \sum_{i=1}^{n} (x_i - E(X))^2 f(x_i)$

$$= \left(-1-\frac{11}{6}\right)^2\cdot\frac{1}{3} + \left(3-\frac{11}{6}\right)^2\cdot\frac{1}{2} + \left(4-\frac{11}{6}\right)^2\cdot\frac{1}{6}$$

$$= \underline{\frac{149}{36} \approx 4,14}$$

c) Mode is the most likely value in X

$\Rightarrow \underline{3}$

d) $C_V = \dfrac{\text{standard deviation}}{\text{Expected value}} = \dfrac{\sqrt{\frac{149}{36}}}{11/6} \approx \underline{1,11}$

2 Frequency analysis and linear regression

a) $P(\geq 1 \geq 100\text{-y flood in } 10 \text{ y period}) = 1 - P(\text{NO} \geq 100\text{y flood in } 10\text{y})$

$P(100\text{-y flood}) = \frac{1}{100}$, $P(\text{Not } 100\text{y flood}) = \frac{99}{100}$

$P = 1 - \left(\frac{99}{100}\right)^{10} \approx \underline{0{,}096 = 9{,}6\ \%}$

b) One assumption that must be fulfilled in simple linear regression is that of homoscedacity.
This means that the error in the regression should be equal for each value.

From plot A it is evident that the error grows for larger values of average runoff, thus the homoscedacity assumption does not hold for these data.

The QQ plot (B) indicates that the data may be normally distributed, with some increase in variance near the extremals - this is usual. (Acceptable normality)

One way to improve the analysis could be only viewing the discharge data between ~ 0-50 m^3/s, as (by visual inspection) these data points seem to fulfill the homoscedacity assumption. One must be careful not to extrapolate the result from a regression for these values to the rest of the data. A regression only holds for the data it is calculated from.

3 Confidence intervals

$n = 30$, $\bar{x} = 145$, var $= 20$

a) 95% confidence on the mean if:

i) Unknown variance $\Rightarrow$ t-distribution w/ $n-1$ dof.

$$s_x = \sqrt{20}, \quad s_{\bar{x}} = \sqrt{\frac{s_x^2}{n}} = \sqrt{\frac{20}{30}} = \frac{\sqrt{6}}{3}$$

Need the table value for $t_{1-\frac{\alpha}{2}, n-1} = t_{0,975, 29}$ as $\alpha = 0,05$.

$$t_t = 2,045$$

$$l = \bar{x} - t_{0,975,29} \cdot s_{\bar{x}} = 145 - 2,045 \cdot \frac{\sqrt{6}}{3} = 143,33$$

$$u = \bar{x} + t_{0,975,29} \cdot s_{\bar{x}} = 145 + 2,045 \cdot \frac{\sqrt{6}}{3} = 146,67$$

95% confidence interval: $\underline{(143,33 \ , \ 146,67)}$

ii) Known variance $\Rightarrow$ normal distribution

$$\sigma_x = \sqrt{20}, \quad \sigma_{\bar{x}} = \frac{\sqrt{6}}{3}$$

$$z_{1-\frac{\alpha}{2}} = z_{0,975} = 1,96$$

$$l = \bar{x} - z_{0,975} \sigma_{\bar{x}} = 145 - 1,96 \cdot \frac{\sqrt{6}}{3} = 143,40$$

$$u = \bar{x} + z_{0,975} \sigma_{\bar{x}} = 145 + 1,96 \cdot \frac{\sqrt{6}}{3} = 146,60$$

95% confidence interval: $\underline{(143,4 \ , \ 146,6)}$

b) When the true variance is unknown and estimated we have a greater uncertainty. This uncertainty is reflected in the wider confidence interval we get from the t-distribution.

3 c) 95% confidence on the variance

$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2} \quad \text{follows a } \chi^2\text{-distribution}$$

$$= \frac{n-1}{\sigma_x^2} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = (n-1)\frac{S_x^2}{\sigma_x^2}$$

$$\Rightarrow \quad L = \frac{(n-1)s_x^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}} = \frac{29 \cdot 20}{\chi_{0{,}975,\, 29}} = \frac{29 \cdot 20}{45{,}722} \approx 12{,}69$$

$$u = \frac{(n-1)S_x^2}{\chi^2_{\frac{\alpha}{2}, n-1}} = \frac{29 \cdot 20}{\chi_{0{,}025,\, 29}} = \frac{29 \cdot 20}{16{,}047} \approx 36{,}14$$

95% confidence interval : $\underline{(12{,}69 \; , \; 36.14)}$

4 Machine learning:

a) The point of ML is to train an algorithm to find a pattern - not more.
Thus - we use the training set to train the algorithm to find a certain pattern in these data.

Then we let the algorithm operate on a fresh test set, and check the error. We want this error to be minimal (this is the test error). The test error can grow in two ways, either due to underfitting or overfitting to the training data.

Underfitting is the case when the algorithm has not been able to discover the pattern of the training set. We will get a large training error - and this will translate into a large test error. The algorithm is not at all trained, e.g. due to too low complexity.

Overfitting is the case when the algorithm has minimised the training error, though, when operating on the test set yields a large test error. This happens because the algorithm has specialised on the training data, i.e. surpassed the level at which it discovers the pattern. This is an indication of an overcomplex algorithm.

We strike for the "Goldilocks" moment of "just right". Meaning, the algorithm has found the pattern in the training data and yields a small test error. Hence, we allow for some training error to get a comparably low test error.

4

b) The complexity governs the overfitting / underfitting balance. By controlling the complexity we can more easily reach the "just right" algorithm. High complexity often results in overfitting, low complexity in underfitting.

If runtime is a concern, then keeping the complexity relatively low will decrease runtime.

5 Time series analysis and Fourier transformation.

a) There are several ways of testing for a significant trend in X_t, e.g. linear regression and Mann-Kendall test.

Here, the run test seems to be suitable for the data.

Start by calculating the mean of the data $\bar{X}$ (could also be the median).

Then we compare each of the data points to the $\bar{X}$-value.

If an $x_i > \bar{X}$, then assign it a +

else assign it −

Define n_1 = number of +

n_2 = number of −

Count the amount of runs, defined as consecutive series of + or −, e.g. R for ++−+−−−+++− is 6

In our case $n_1, n_2 > 10$ => the algorithm states that R is approximated by a normal distribution w/

mean: $\mu = \frac{2 \cdot n_1 \cdot n_2}{n_1 + n_2} + 1$

variance: $\sigma^2 = \frac{2 \cdot n_1 \cdot n_2 \cdot (2 \cdot n_1 \cdot n_2 - n_1 - n_2)}{(n_1 + n_2)^2 \cdot (n_1 + n_2 - 1)}$

To check the trend, calculate the test statistic

$$Z = \frac{(R - \mu)}{\sigma}$$

If $|Z| > z_{1-\frac{\alpha}{2}}$ we have a significant trend, with α being the significance level of choice.

5

b) From the X_t-plot, it is evident that the data are periodic, with a main periodicity of ~5 seconds

In frequency space - this should correspond to a significant peak at $f = \frac{1}{T} = \frac{1}{5} = 0{,}2\ ^1/_s$

Plot A shows just this. As the values of X_t are purely real, the data set X_k is symmetric about ½. Therefore, we also get a peak at 0,8.

Plot A shows the Fourier transform of X_t.