Exam GEO4300
23.11.2020
Cand.nr: 15402

1 – Random variable parameter estimation

Discrete random variable is defined as

$$X = \begin{cases} -1, & p = \frac{1}{3} \\ 3, & p = \frac{1}{2} \\ 4, & p = \frac{1}{6} \end{cases}$$

a) Find the expected value

$$E(X) = -1 * \frac{1}{3} + 3 * \frac{1}{2} + 4 * \frac{1}{6} = \frac{11}{6}$$

b) Find the variance

$$V(X) = \sigma^2 = E[(X - E(X))^2] = E(X^2) - E^2(X)$$

$$E(X^2) = (-1)^2 * \frac{1}{3} + 3^2 * \frac{1}{2} + 4^2 * \frac{1}{6} = \frac{15}{2}$$

$$= \frac{15}{2} - \frac{11}{6}^2 = \frac{149}{36} \approx 4.14$$

c) Find the mode

The mode is 3, since the probability for this value is the highest of the possible probabilities.

$$f(3) = \frac{1}{2}$$

d) Find the coefficient of variation

$$Cv = \frac{\sqrt{V(X)}}{E(X)} = \frac{\sqrt{\frac{149}{36}}}{\frac{11}{6}} = 1.1$$

2 – Frequency analysis and linear regression

a) *Probability for at least one 100-year flood in a period of 10 years?*

Probability of 100-year flood: p=0.01
Return period: T=100

$$\binom{10}{1}(0.01)^1(0.99)^9 = 0.09$$

b) *Which assumptions of linear regression is violated and what strategies can improve the analysis?*

In the plots of this task we have a heavy tailed QQ-plot and a regression which show a line almost at y=x. To get this kind of plot there has been assumed normality in the regression. This means one of the distributions in the scatter plot is normally distributed while the other is not. Since the QQ plot curve at the ends the distribution probably has more extreme values than expected.

To improve the analysis one can try to remove the extreme values from the datasets.

3 – Confidence intervals

Sample with 30 random observations, produced mean of 145 and variance of 20.

a) *What is the 95% confidence interval on the mean assuming a normal distribution if:*
   a. *The true variance is unknown and estimated to be 20?*

If the true variance is unknown, we use the t-distribution to estimate the confidence interval on the mean:

$$s_{\bar{x}} = \frac{\sqrt{var}}{\sqrt{n}} = \frac{\sqrt{20}}{\sqrt{30}} = 0.82$$

From table values for t, we get:

$$t_{1-\frac{\alpha}{2},n-1} = t_{1-\frac{0.05}{2},29-1} = t_{0.975,28} = 2.048$$

The confidence intervals are then:

$$l = \bar{x} - t_{1-\frac{\alpha}{2},n-1} * s_{\bar{x}} = 145 - 2.048 * 0.82 = 143.3$$
$$u = \bar{x} + t_{1-\frac{\alpha}{2},n-1} * s_{\bar{x}} = 145 + 2.048 * 0.82 = 146.7$$

The true mean lies within the interval from 143.3 to 146.7.

   b. *The true variance is 20?*

If the true variance is known, we can use the z-distribution to estimate the confidence intervals:

$$\sigma_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{\sqrt{20}}{\sqrt{30}} = 0.82$$

From table values for z, we get:

$$z_{1-\frac{\alpha}{2}} = z_{1-\frac{0.05}{2}} = z_{0.975} = 1.96$$

The confidence intervals are then:

$$l = \bar{x} - z_{1-\frac{\alpha}{2}} * \sigma_{\bar{x}} = 145 - 1.96 * 0.82 = 143.4$$
$$u = \bar{x} + z_{1-\frac{\alpha}{2}} * \sigma_{\bar{x}} = 145 + 1.96 * 0.82 = 146.6$$

The true mean lies within the interval from 143.4 to 146.6.

b) What is the reason for the different results?

We see the confidence interval calculated when knowing the true variance is slightly smaller than the interval calculated when the true mean was unknown. This difference comes from the use of two different probability distribution, the t-distribution and the z-distribution.

The z-distribution is the standard normal distribution, while the t-distribution is an approximation of the normal distribution. The normal distribution is more peaked and has shorter tails than the t-distribution. Since we use the t-distribution with 1 degree of freedom, this difference is more pronounced than if we had used a higher degree of freedom.

Since our sample size is 30, the t-distribution is almost identical to the normal distribution and can be approximated to it. That is why the number are different, but only slightly.

c) What is the 95% confidence interval on the variance?

To find the 95% confidence interval on the variance we use values from the Chi-squared distribution.

$$\chi^2_{\frac{\alpha}{2},n-1} = \chi^2_{0.025,28} = 44.461$$
$$\chi^2_{1-\frac{\alpha}{2},n-1} = \chi^2_{0.975,28} = 15.308$$

$$l = \frac{(n-1){s_x}^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} = \frac{28 * 20}{15.308} = 36.6$$
$$u = \frac{(n-1){s_x}^2}{\chi^2_{\frac{\alpha}{2},n-1}} = \frac{28 * 20}{44.461} = 12.6$$

The true variance lies within the interval 12.6 to 36.6.

4 – Machine learning

a) Why is it common to split the dataset into training set and test set when doing machine learning?

It is common to split the dataset to avoid the learning to be biased of the data you train the model with. The model is a result of the data you train it with, and if you test the model with the same data you fit it to, it will always be a good fit. When the model is then tested on other data, it is sure to perform poorly. An extreme case of this is overfitting, where the model is too well adjusted to the training data is will always lead to bad results on other data.

b) In many machine learning algorithms, you have a parameter which controls the complexity of the model. Why do we want to control this complexity?

We want to control the complexity of the model to be sure it will perform as well as we want it to. If the model is not complex enough, it will be underfitted and show a generalization of the data instead of what we want it to show. If the model is too complex, we have overfitted it, making it so complex it will not give good results on other data. In the overfitting case the test error will be large, but the training error will likely be very small. If the model is underfitted both training error and test error will be large.

To best control the complexity of the model we must find the complexity which gives the smallest test error. The data must be split into training set and test set, and the model must only be tested on the test set. When we have found the smallest test error, we can validate the training sets by cross validation or bootstrapping.

5 – Time series analysis

a) How could you test if there is a significant trend in the time series X_t?

The time series X_t is a continuous measurement series where a parameter is measured every second. We have 30 measuring steps. A suitable trend test for this time series is either the 'run test'-method, the Mann-Kendall trend test or a linear regression trend test.

The 'run test'-method is a simple method to find a trend in a time series. First the mean $\bar{X}$of the time series is calculated. Then each value X_i in the time series is compared to the mean to see if it is higher or lower. If it is higher, the value is given a '+'-sign, if it is lower it is given a '-'- sign. Then the number of 'runs' is calculated. A run is when multiple '+' or '-' values show up in a row. The number of '+'-rows and '-'-rows are split into two groups n_1 and n_2. If both of them are higher than 10 the distribution is approximated by a normal distribution with mean:

$$\mu = \frac{2 * n_1 * n_2}{n_1 + n_2} + 1$$

and variance:

$$\sigma^2 = \frac{2n_1n_2 * (2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

To test for trends the Z-statistic is used:

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

where R is the total number of runs, $R=n_1+n_2$.

If the absolute value of the test statistic is higher than the calculated table value for a certain significance value, there is a trend.

$$|Z| > Z_{1-\frac{\alpha}{2}}$$

b) Which of the graphs A, B or C show the Fourier transformation of X_t?

Figure C show the Fourier transformation of X_t.