Exam GEO4300 15414

November 23, 2020

1 Random variable parameter estimation a) The expected value is:

$$E(X) = \sum x_j f(x_j) = -1 * 1/3 + 3 * 1/2 + 4 * 1/6 = 11/6$$

b) The variance is:

$$\sigma_x^2 = \sum (x_j - \mu)^2 * p(x_j) = 4.14$$

where μ is 11/6 p are the probabilities.

c)The mode is the most frequently occuring value. 3 has the highest probability and is thus the mode value.

d) The coefficient of variation is the ratio of the ratio between the standard deviation and the expected value. The standard variation we find by simply taking the square root of the variance.

$$Cv = \frac{\sqrt{V(x)}}{E(x)} = 2.26$$

0.1 2. frequency analysis

a) Using the binomial distribution the probability of one flood exceeding the 100 year flood in 10 years is:

$$f_x(x; n, p) = \frac{10!}{1!(10-1)!} 0.01 * (1 - 0.01)^2 = 0.098$$

b) An assumption of linear regression is that the data is normally distibuted. we can see that the tails of the QQ plot start do diverge from the straight line. If the data was normally distributed all the points should fall more or less on the line, so the assumption of normality is violated. This is likely generated by the outliers we see in the plot, they could for example be caused by 100 year floods. A method for improving this analysis ould be to remove extreme events, like 100 year floods and alanyse them seperately using extreme vlue distributions.

0.2 3 Confidence intervals

a)

i)We have a small amount of observations, therfore we assume that the true variance is unknown. We thefrore use the t distribution with n-1 degrees of freedom wich gives us the 95% confidence interval:

$t_{1-\alpha/2,n-1S_{\bar{x}}} = 2.045$

$S_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{\sqrt{20}}{\sqrt{30}} = 0.817$

$$l = \bar{x} - t_{1-\alpha/2,n-1S_{\bar{x}}} = 135.85$$

$$u = \bar{x} + t_{1-\alpha/2,n-1S_{\bar{x}}} = 154.15$$

ii) With an known variance we use the standard normal distribution: We use the same standard deviation as above and get the Z value

$$z_{0.975} = 1.96$$

$$l = \bar{x} - z_{0.975}\sigma_{\bar{x}} = 136.24$$

$$u = \bar{x} + z_{0.975}\sigma_{\bar{x}} = 153.77$$

b) The t distribution generally has a higher kurtosis than the normal distribution,meaning that its tail are larger. This gives it a higher probability of having values far away from the mean. This should give a larger CI for the t distribution which is what we see in i and ii.

c) If our observations are normally distributed we can use a Chi-squared distribution to find the CI for the variances:

$$\chi^2_{\alpha/2,n-1} = 45.72$$

$$\chi^2_{1-\alpha/2,n-1} = 16.05$$

$$l = \frac{(n-1)s_x^2}{\chi^2_{1-\alpha/2,n-1}} = 12.69$$

$$u = \frac{(n-1)s_x^2}{\chi^2_{\alpha/2,n-1}} = 36.14$$

0.3 4 Machine learning

a) When doing machine learning it is common to split the data into test and training sets in order to avoid overfitting your data. The machine learning algorithm will to some degree be biased towards the training data, testing it on an independent set of data will serve to expose to what degree the algorithm is biased. The model may for example have a very low training error, but high test error. This means that the model very closely fits the data it was trained on, but when generalized to an independent test set is fails due to overfitting.

b) It is essential to control the complexity of a model since you can fit any number of parameters to a training dataset. You can even include more parameters than there are datapoints. This will leave you with a model that very accurately describes your training dataset, but wich will generalize very badly. If your training set includes a lot of random noise, fitting to many parameters will cause the model to start fitting to the noise as well, wich is not desirable in most cases. To much complexity also comes at a computational cost, meaning longer runtimes.

0.4 5 time series analysis and Fourier transformation

a) In order to test if there is a significant trend in the data we can use a linear regression method with a t test to see if the trend is significantly different from 0. First we fit a linear line to the data with time as the independant variable. We then want to see if the fitted line has a slope significantly different from zero. If it is significantly different from zero, there is a trend. The t test has the test statistic:

$$t = \frac{\beta - 0}{S_\beta}$$

where β are the model coefficients and S_β is the standard deviation of these coefficients.

The hypothesis that there is no trend is rejected if

$$|t| > t_{1-\alpha/2,n-2}$$

b) In the graph of $X(t)$ we see a clear signal with a frequency of abut $1/5$ Hz. This is the most pronounced signal and should therfore show up as the largest peak in the fourier transform. The signal also has some additional noise making it slightly irregular, wich should give some additional peaks in the frequency domain. This leads me to conclude that graph A shows the fourier transform of X_t.