Exam in GEO4300/GEO9300 - 2020

In [2]:
```
import numpy as np
import scipy.stats as st
```

1. Random variable parameter estimation

(a)

The expected value can be calculated as:

$$E(X) = (-1) \cdot \frac{1}{3} + 3 \cdot \frac{1}{2} + 4 \cdot \frac{1}{6} = \frac{11}{6} \approx 1.83$$

(b)

The variance is given by

$$V(X) = E(X^2) - (E(X))^2$$

so we calculate

$$E(X^2) = (-1)^2 \cdot \frac{1}{3} + 3^2 \cdot \frac{1}{2} + 4^2 \cdot \frac{1}{6} = \frac{15}{2} = 7.5$$

and

$$(E(X))^2 = \left(\frac{11}{6}\right)^2 = \frac{121}{36} \approx 3.36$$

Pluggin these values into the equation for the variance, we get

$$V(X) = \frac{15}{2} - \frac{121}{36} = \frac{149}{36} \approx 4.14$$

(c)

The mode is given by the number with most occurences, and is in this case 3

(d)

The coefficient of variance is given by the ratio of the standard deviation and the mean. The standard deviation is given by $\sigma = \sqrt{V(X)}$, thus the equation for the coefficient of variance becomes

$$\frac{\sqrt{V(X)}}{E(X)} = \frac{\sqrt{4.14}}{1.83} \approx 1.11$$

2. Frequency analysis and linear regression

(a)

To calculate the probability to observe at least one 100-year flood or larger within a period of 10 years, we first need to calculate the probability of not observing any 100-year flood within the given period. This is done by

$$(1 - 1/100)^{10} \approx 0.9044$$

Then we subtract that number from one, giving us the probability to observe at least one 100-year flood within the given beriod.

$$1 - 0.9044 = 0.0956$$

And so, the probability to observe at least one 100-year flood is 9.56 %

(b)

The assumptions we make for a simple linear regression is:

- Homoscedacity
- Normality
- Linearity
- Independence

Linearity is assessed from the correlation, as the correlation is a measurment of the linear relationship between two variables. Thus, we use the scatterplot. From the scatterplot in 1A we can see that there is an indication of positive linearity, but there is also some outliers that preferably should be removed. To assess this more certainly, it would an idea to calculate the pearson correlation value itself. Anything betwen |0.5| and |0.7| would be a moderate correlation and anything above |0.7| would be accepted as good correlation.

From the Q-Q plot in 1B we observe that the residuals are not normally distributed. For it to show that the residuals are normally distributed, the points have to align in a diagonal from bottom left to upper right. The Q-Q plot showed in this exercise, is what is called thick tailed, and suggests that the data contains residuals with extremely high values compared to what would be expected of a normal distribution. To achieve normal distribution, we would suggest taking the logarithm of the data.

The homoscedacity is assessed by a scatter plot of the residuals, and the independence is assessed through autocorrelation, both of which we do not have information about.

3. Confidence intervals

(a)

the 95 % confidence interval on the mean assuming

(i) true variance is known and estimated as 20 ($s_x = \sqrt{20}$)

Assuming normal distribution, we can use a t-distribution to calculate the lower and upper limits of the confidence interval

$$L = \bar{x} - t_{1-\alpha/2,n-1} s_{\bar{x}}$$
$$U = \bar{x} + t_{1-\alpha/2,n-1} s_{\bar{x}}$$

the t-value can be calculated using

In [4]:
```
alpha = 0.05
n = 30
t = st.t.ppf(1-alpha/2, n-1)
print('t-value:', t)
```

```
t-value: 2.045229642132703
```

with $\bar{x} = 145$ and $s_{\bar{x}} = s_x/\sqrt{30} \approx 0.816$ we get

$$L = 145 - 2.045 \cdot 0.816 \approx 143.33$$
$$U = 145 + 2.045 \cdot 0.816 \approx 146.67$$

And thus, the 95 % confidence interval about the mean assuming unknown true variance is given by [143.33, 146.67]

(ii) true variance is 20 ($\sigma_x = \sqrt{20}$)

For this, we can use the standard normal distribution to calculate the confidence interval.

$$L = \bar{x} - z_{1-\alpha/2}\sigma_{\bar{x}}$$
$$U = \bar{x} + z_{1-\alpha/2}\sigma_{\bar{x}}$$

we calculate the z-value by using python

In [6]:
```
z = st.norm.ppf(1-alpha/2)
print('z-value:', z)
```

```
z-value: 1.959963984540054
```

Again, we have that $\bar{x} = 145$ and $\sigma_{\bar{x}} = \sigma_x/\sqrt{30} \approx 0.816$. Plugging these values into the equation, we get

$$L = 145 - 1.960 \cdot 0.816 \approx 143.40$$
$$U = 145 + 1.960 \cdot 0.816 \approx 146.60$$

The 95 % confidence interval about the mean with known true ariance is given by [143.40, 146.60]

(b)

The difference is explained by using the t-distribution for when the true variance was unknown, whilst the z-distribution was used when the true variance was known. When the number of observations increase, then the difference in values we get by using either t-distribution or z-distribution will decrease. In this case, we had 30 observations, which is on the border of when it is okay to use z-distribution. Had the observations been smaller, then it would be much more preferable to use t-distribution.

(c)

The limits for the 95 % confidence interval on the variance is given by

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

where chi squared can by calculated using python

In [9]:
```
chisq1 = st.chi2.ppf(1-alpha/2, n-1)
chisq2 = st.chi2.ppf(alpha/2, n-1)
print('chi squared for lower:', chisq1)
print('chi squared for upper:', chisq2)
```

```
chi squared for lower: 45.72228580417452
chi squared for upper: 16.04707169536489
```

with the estimated variance $s_x^2 = 20$ and $n = 30$, we can find the lower and upper limits

$$L = \frac{(30-1)20}{45.722} = 12.68$$
$$U = \frac{(30-1)20}{16.047} = 36.14$$

The 95 % confidence interval on the variance is given by [12.68, 36.14]

4. Machine Learning

(a)

When doing machine learning, we choose a variable Y that we want to predict, and variables ($X_1, X_2, X_3, \ldots$) that we want to use as predictors for $Y 4. Further on, we collect the data that we want, and remove any outliers or missing data. Now, we can split the data into a training set and a test set. The preferable ratio would be to have 70% of the data in the training set, and 30 % of the data in the test set. We can perform various machine learning models on the training set, and later use the test set to test the machine learning models we used on the training set to assess how the model works on the data. The reason we split into two different sets is because the training error (the prediction error of the training set) overestimates the test error (the prediction error of the test set), meaning that it would be too large uncertainties if we only had one set that both trained and tested the model.

(b)

If the model is not complex enough, then it will underfit the data. This is a bad generalization, and leads to large training error and test errors. If the model is too complex, it will overfit the data. Again, this is bad for generalization and leads to small training errors, but large test errors. Usually, the training set will always prefer more complexity, but we want to control this so that we do not overfit. The preferable outcome would be to have small test errors.

5. Time series analysis and Fourier transformation

(a)

A suitable test would be to use a linear regression model for a confidence level of $\alpha = 10$. Firstly, use the `statsmodels.formula.api.ols(formula = 'X_t ~ T).fit()'` function in python (which gives you the ordinary least squares). From this, one can extract the coefficient for the intersect and the independent variable T (β), which then allows you to make a simple linear regression model. Further on, we can do a t-test to check if the trend (the coefficient to the independent variable) is significantly different from zero. To conduct such a test, we first have to state a null hypothesis saying that the trend is not significant. The alternative hypothesis is of course that the trend *is* significant. The null hypothesis is rejected if $t_{stat} > t_{1-\alpha,n-2}$, where α is already stated and n is the number of X_t values. The t_{stat} can be caluclated as $t_{stat} = \beta / s_\beta$, where s_β is the standard deviation of coefficient β, and can be read from the summary of the ordinary least square model.

(b)

As a plots of Fourier transforms are a plot of the magnitude, and we can thus find the highest peak by simply looking at the frequency of the peaks of X_t. We observe that the peaks occure with a frequency of 5 seconds, and so we can find where we expect the peak in the Fourier transform to be by simply 1/5 = 0.2. As figure A is the only one showing a its highest peak at 0.2, then this is the figure showing the Fourier transform of X_t. The other peak we see in A at 0.8 is due to mirroring of the 0.2 frequency at the other side of the Nyquist frequency.

In [ ]: