Exam GEO4300 23.11.20

Candidate: 15412

In [1]:

```
import scipy.stats as st
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Question 1

a)

The expected value is given by $\mu = E[X] = \sum_i x_i f(x_i)$.

In [2]:

```
E=-1*(1/3)+3*0.5+(4*(1/6))
print('The expected value is: %.2f.' %E)
```

```
The expected value is: 1.83.
```

b)

The variance is given by $\sigma^2 = V[X] = E[(x-\mu)^2] = \sum_i (x_i-\mu)^2 f(x_i)$.

In [3]:

```
V=(-1-1.83)**2*(1/3)+(3-1.83)**2*(1/2)+(4-1.83)**2*(1/6)

print('The variance value is: %.2f.' %V)
```

```
The variance value is: 4.14.
```

c)

The mode is the most frequently occurring value so in for this variable X the mode is 3.

d)

The coefficient of variation is given by $CV = \sigma/\mu$

```
In [4]: Cv=(V**(1/2))/E
        print('The coefficient of variation is: %.2f' %Cv )
```

```
The coefficient of variation is: 1.11
```

Question 2

a)

```
In [5]: P=1-((99/100)**10)
        print('The probability to observe at least one')
        print('100-years flood or larger within a period of 10 years is: %.4f' %P)
```

```
The probability to observe at least one
100-years flood or larger within a period of 10 years is: 0.0956
```

b)

The asumption of homoscedasticity is violated in these figures. This is because the variance of residual is not the same for any value. For the larger values the spread is much bigger than for the smaller values. To improve the anlysis we can revove the extreme values so we only have a linear regresion for the values where the Normal Q-Q plot is linear.

Question 3

a)

In [6]:

```
#i)
n=30
mean=145
var=20
alfa=0.05
s=var**(1/2)

s_x=s/(n**(1/2))

t_est30=st.t.ppf(1-alfa/2,n-1)
l_est30=mean-t_est30*s_x
u_est30=mean+t_est30*s_x
print('The 95 % confidence interval for the mean with estimated var
iance is:')
print(l_est30,u_est30)
```

```
The 95 % confidence interval for the mean with estimated variance
is:
143.33007699 146.66992301
```

In [7]:

```
#ii)
t_true30=st.norm.ppf(1-alfa/2)
l_true30=mean-t_true30*s_x
u_true30=mean+t_true30*s_x
print('The 95 % confidence interval for the mean with true variance
is:')
print(l_true30,u_true30)
```

```
The 95 % confidence interval for the mean with true variance is:
143.399696108 146.600303892
```

b)

The reason for the differences in i) and ii), are that we use the t-value in i) and the z-value in ii). The t-value uses a estimated variance where we estimate the variance from a sample mean and in order to have an unbiased estimation we use n-1 for number of values in the sample. In the z-value on the other hand we use the true variance, found by using the population mean, and we uses n values in the population. The error in i) is therfore bigger than ii).

c)

```
In [8]: l_chi=((n-1)*var)/st.chi2.ppf(1-alfa/2,n-1)
        u_chi=((n-1)*var)/st.chi2.ppf(alfa/2,n-1)
        print('The 90 % confidence interval for the variance is:')
        print(l_chi,u_chi)
```

```
The 90 % confidence interval for the variance is:
12.685280051 36.1436660227
```

Question 4

a)

It is common to split the dataset into a training set and a test set. This is because the model is often made frome the input data comming from a single dataset, and the model should also predict other observations that was not in the dataast as well. It is better to estimate error from other data than the one used to make the learner. This is becouse the training error (the prediction error of the training set) often underestimates the test error (the prediction error of the test set). The split is used to stop model overfitting, where the model is to fitted to the training set, and that makes the model worse to make predictions for new data.

b)

It is importent to control Model Complexity, to make the model have a suitable fit. If we make the model too complex it gets overfitted.The train error, gets small, but the test error gets too large. If the model is too simple on the other hand the model is underfitted with large test and train errors. We therfore need to find the model clomplexity with the smallest test error. To find thisthe test set should not be used. Insetead we can use a validation set from the training set to do Bootstrap or cross-validation.

Question 5

a)

To test if there is a significant trend in X_t i would use a ordinary least squares test (OLS). The test is used to estimate the unknown parameters in a linear regresion model. The OLS uses the parameters from the linear regresion function by finding the smallest sum of the squares in the differenses between the observed values and the values from the linear function.

b)

Figure A shows the Fourier transform of X_t . This is because we can see in X_t that a peek is every 5 secounds. So the freqency peek in the Fourier plot should be at 1/5s= 0.2 [1/s]. This peak can we see in A.