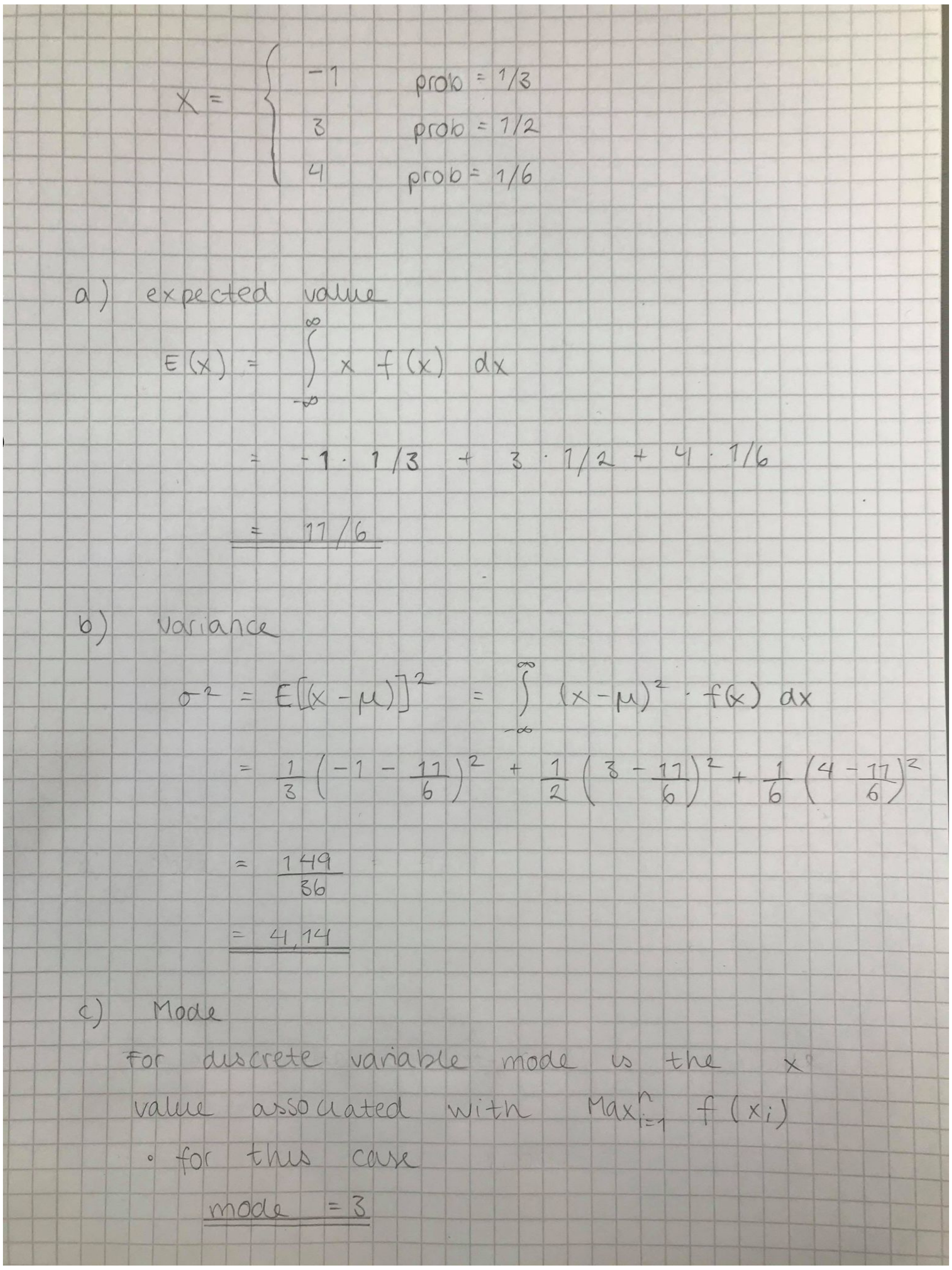
Exam GEO4300

23. November 2020

Candidat nr. 15415

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as st
```

Problem 1 Random variable Parameter estimation



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

a) expected value

$$E(x) = \int_{-\infty}^{\infty} x \; f(x) \; dx$$

$$= -1 \cdot 1/3 + 3 \cdot 1/2 + 4 \cdot 1/6$$

$$\underline{= 11/6}$$

b) variance

$$\sigma^2 = E[(x-\mu)]^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x) \; dx$$

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

$$= \frac{149}{36}$$

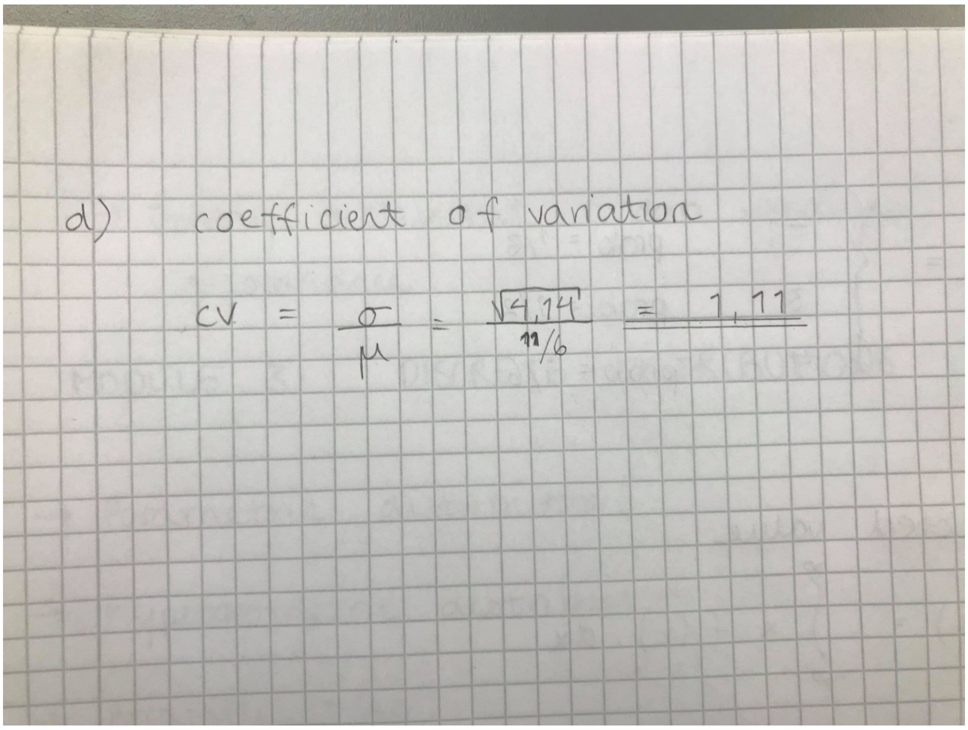
$$\underline{\underline{= 4,14}}$$

c) Mode

For discrete variable mode is the x value associated with $\text{Max}_{i=1}^{n} f(x_i)$

· for this case

$\underline{\text{mode} = 3}$



d) coefficient of variation

$$CV = \frac{\sigma}{\mu} = \frac{\sqrt{4,14}}{11/6} = \underline{\underline{1,11}}$$

Problem 2 Frequency analysis and linear regression

The probability to observe at least one 100-years flood or larger with a period of 10 years is 1 minus the probability that it will not happend during the period

```
In [3]: #a )
        P = 1- (1-(1/100))**10
        print("The probability to observe at least one 100-years flood or larg
```

```
The probability to observe at least one 100-years flood or larger wi
th a period of 10 years is 0.096
```

b)

Assumptions simple linear regression:

· Linearity

· Independence

· Homoscedasticity

· Normality

Homoscedasticity is violated in this analysis.

Problem 3 Confidence interval

In [8]:
```
alpha = 0.05
n = 30
x = 145
var = 20
std = np.sqrt(var)
```

a)

i

· unknown variance

· assume normal distributed

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

In [9]:
```
std_x = std/(np.sqrt(n))
t = st.t.ppf(1- alpha/2, n-1)

L = x - t*std_x
U = x + t*std_x

print("If the estimated variance is 20, the 95% confidence interval on
```

```
If the estimated variance is 20, the 95% confidence interval on the
mean is 143.33 to 146.67
```

ii

· Standard normal distributed

· Known varianse

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

In [12]:
```
z = st.norm.ppf(1- alpha/2)

sigma_x = std/(np.sqrt(n))

L1 = x - z*sigma_x
U1 = x + z*sigma_x

print("If the estimated variance is 20, the 95% confidence interval on
```

```
If the estimated variance is 20, the 95% confidence interval on the
mean is 143.4 to 146.6
```

b) There is a small diffence between the confidence interval when the variance is estimated and known. The confidence interval is slightly smaller when the varince is known. This difference is a result by a bigger uncertainy when the variance is estimated and not known. In the calculation for confidense interval with an estimated variance we have n samples. When the variance is known, n is much bigger.

c)

chi-square distribution

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

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

In [13]:

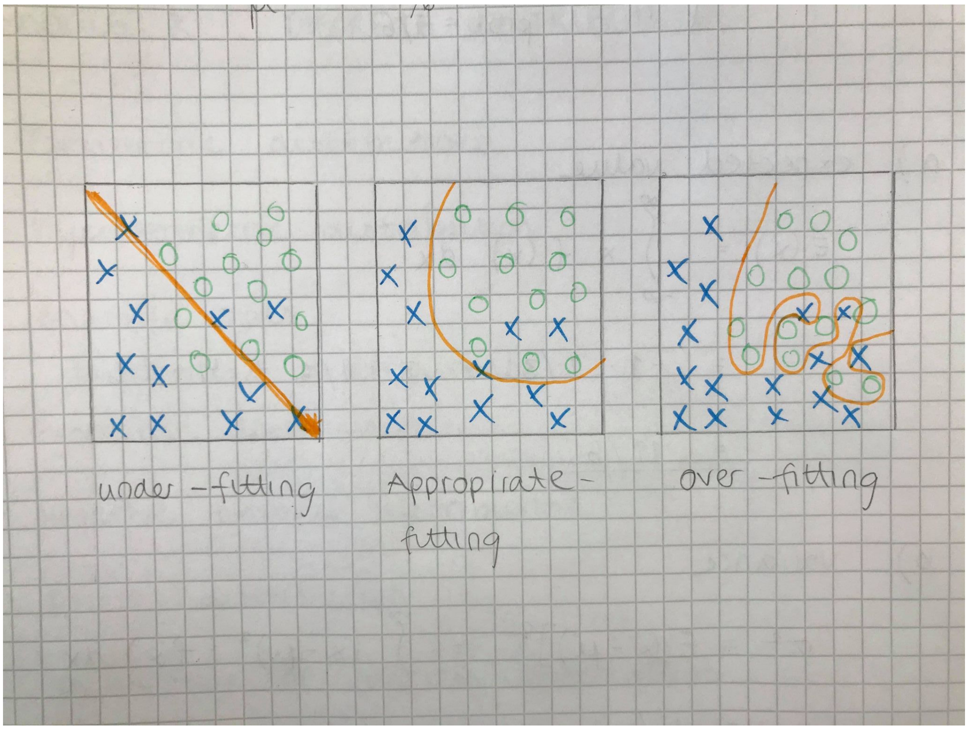
```
x1 = st.chi2.ppf(1-alpha/2, n-1)
x2 = st.chi2.ppf(alpha/2, n-1)

L2 = ((n-1)*var)/x1
U2 = ((n-1)*var)/x2

print("The 95% confidence interval for the variance is", np.round(L2,2
```

```
The 95% confidence interval for the variance is 12.69 to 36.14
```

Problem 4 Machine learning

a) In meachine learning we want to create an alogrithm that finds a pattern in our data. To do that, we split our data set in two parts (a training set and a test set). The training set contains 2/3, and the test set contains 1/3 of our data points, both part must to be representative of the data. We use the training set to teach it how to find the results we want (find the pattern). While doing this, the test set is hidden. After teaching the training set, we use the alogrith on the test set. It is important to be aware that the test-error and training error is not necassery the same. Training error is the error we get when we use the alogrithm on the training data again, and the test error is the error we get when we use the trained algorithm on the test data that is has never been exposed to. Error is further described in the next part-question.

b) Typically, a machine learning algorithm include a parameter adjusting the complexity of your model. Complexity is a number of features we use in the training data to predict our outcome. If this complexity is too simple, the algorithm is underfitting. This results in a large train error, and a large test error (illustrated in figure below). If the complxity is too complex, the algorithm is overfitting. This results in a small train error and a large test error (illustrated in figure below). We want to find the "Sweet spot", where the model complexity that gives smallest test error.



Problem 5 Time-series analysis and Fourier transformation

a) Ordinary least square method

To test if there is a significant trend in X_t , I would use a ordinary least square method to find the necessary variable to make a linear regression. The goal in linear regression is to make a function that fit the data. This function that is made is a line ($Y = \alpha + \beta T$) through the data point.

Further, I would use a linear regression method to test if the trend is significant (The slope β)

$$H_0 = \text{No trend}, \beta = 0 \quad H_a = \text{Significant trend}, \beta \neq 0$$

Test statstic:

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

$S_\beta = \frac{S}{\sqrt{\sum (T_i - \bar{T}_i)^2}}$, and $S = \sqrt{\frac{1}{n-2} \sum (Y_i - \hat{Y}_i)^2}$

S_β is the standard deviation og the coefficient β, and S is the standard deviation of the regression. Y_i and $\bar{T}_i$ are observed and estimated hydrologic variables.

The null hypothesis is rejected if $|t| > t_{1-\alpha/2, n-2}$

b) Graph A shows the Fourier transformation of X_t .

The time series X_t are samples once per sencond. The graph illustrate peaks in the timeseries with a timeperiod of 5 seconds (peak every 5 second). The Fourier transformation of X_t is given in frequency [1/s]. With peaks every 5 second, we have a frequency of 0.2 [1/s]. Graph A is thereforethe Fourier transformation of X_t , where we find a peak at 0.2. Graph A is symetric around 0.5, resulting in a peak at 0.8.

```
In [ ]:
```