Exam GEO4300

23/11/2020

In [36]:
```
import numpy as np
import pandas as pd
import scipy.stats as st
```

1 Random variable parameter estimation

A discrete random variable X is defined by

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

(a) Find the expected value

$$E(X) = \sum x_j \cdot f(x_j)$$

In [2]:
```
E = np.sum((-1*1/3) + 3*1/2 + 4*1/6)
print('The expected value E(X): %.2f' % E)
```

```
The expected value E(X): 1.83
```

(b) Find the variance

$$V(X) = E(X^2) - E^2(X)$$
$$E(X^2) = ((-1)^2 \cdot 1/3) + (3^2 \cdot 1/2) + (4^2 \cdot 1/6) = 7.5$$
$$E^2(X) = 10^2 = 3.36$$

In [3]:
```
EX2 = np.sum((((-1)**2)*1/3) + ((3**2)*1/2) + ((4**2)*1/6))
E2X = E**2

V = EX2 - E2X
print('The variance V(x): %.2f' % V)
```

```
The variance V(x): 4.14
```

(c) Find the mode

The mode is the value that appears most in a set of data values, and in the given function 3 will be the mode because $p = 1/2$.

(d) Find the coefficient of variation

$$C_V = \frac{\sigma}{\overline{x}} = \frac{\sqrt{V(X)}}{E[X]} = \frac{\sqrt{4.14}}{1.83}$$

In [13]:
```
CV = np.sqrt(V)/E

print('The coefficient of variation CV(X): %.2f' % CV)
```

```
The coefficient of variation CV(X): 1.11
```

2 Frequency analysis and linear regression

(a) What is the probability to observe at least one 100-years flood or larger within a period of 10 years?

T = 100

n = 10

$$P = 1 - (1 - \frac{1}{T})^n = 1 - (1 - \frac{1}{100})^{10}$$

In [35]:
```
T = 100
n = 10

P = 1 - (1 - (1/T))**n
print('The probability is %.3f' % P)
```

```
The probability is 0.096
```

(b) Figure 1A shows a simple linear regression between average runoff and median annual flood. Figure 1B shows the QQ-plot of the residual where the theoretical quantiles were calculated using the normal distribution. Describe which assumption of a simple linear regression is violated in this analysis, and discuss strategies that can be used to improve the analysis.

The assumption that the dataset is normality is violating in this analysis, we can see in Figure 3a) that the largest values do not coincides on the line and are thus difficult to predict. In figure 3b) there is a QQ-plot, this plot tells us whether the quantiles from the theoretical distribution correspond to the distribution we have assumed (which is the normal distribution in this situation). We see that the quantiles of the most extreme events do not coincide, and that a normal distribution may not be the most appropriate distribution. We can use other methods and distributions to improve this analysis, methods that take into account extreme events, e.g. Gumbel distribution, Pearson III distribution, Log-Pearson III distribution and log-normal distribution. You may also want to use all of these methods and then choose a method that is not too conservative or the opposite.

3 Confidence intervals

A sample of 30 random observations produced a mean of 145 and variance of 20.

(a) What is the 95% confidence interval on the mean assuming a normal distribution if

- **the true variance is unknown and estimated as 20**
- **the true variance is 20**

In [31]:
```
n = 30
mean = 145
var = 20
alpha = 0.05

# UNKOWEN VARIANCE: t-test
# Find the standard deviation
std1 = np.sqrt(var/n)

# Find t-value since unkowen variance
t_value = st.t.ppf(1-(alpha/2), n-1)

# Lowe and upper CI
l1 = mean - t_value*std1
u1 = mean + t_value*std1


# KNOWN VARIANCE: Z-test
# Find the standard deviation
std2 = np.sqrt(var/n)

# Find Z-value since unkowen variance
Z_value = st.norm.ppf(1-(alpha/2))

# Lowe and upper CI
l2 = mean - Z_value*std2
u2 = mean + Z_value*std2

print('The confidence interval on the mean and an estimated varianc
e: (%.2f, %.2f)' % (l1, u1))
print('The confidence interval on the mean and a true variance: (%.
2f, %.2f)' % (l2, u2))
```

```
The confidence interval on the mean and an estimated variance: (14
3.33, 146.67)
The confidence interval on the mean and a true variance: (143.40,
146.60)
```

(b) What is the reason the difference in results in part (1) and part (2)?

When we know the true variance and it is no longer estimated, we are more certain about our estimate and can therefore use the normal distribution that are more narrow than t-distribution and gives a more narrow confidence interval.

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

In [32]:
```
# CI FOR VARIANCE: CHI-TEST
upper_chi = st.chi2.ppf(1-alpha/2, n-1)
lower_chi = st.chi2.ppf(alpha/2, n-1)

# The 95%-confidence interval on the variance
l3 = ((n-1)*var)/upper_chi
u3 = ((n-1)*var)/lower_chi

# Result
print('The confidence interval on the variance: (%.2f, %.2f)' % (l3
, u3))
```

```
The confidence interval on the variance: (12.69, 36.14)
```

4 Machine learning

(a) Why is it common to split the dataset into a training set and a test set when doing machine learning? In your answer, include in a relevant way the terms training error and test error

Splitting the dataset into training and test sets is an important part of evaluating models in machine learning. We want to train the models and use it on an unknown data set to evaluate models, this also means that you can understand the models better. If not then you can get training errors when you run the trained model back on the training set, and you will not see how good or bad the models predict. While test error is the error when the trained model is running on an unknown data set.

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

We want to check the complexity of a model and determine the number of features we want to include in the model so that we do not overfit the model for the training set. Because if we do that, we will get bad generalization of the model and we will get high test error but low training error since we have adapted the model to the training set. However, if the model is too simple, it will also lead to poor generalization, and will result in both major test and training errors. We want to find the sweet spot, where the model gives smallest test error.

5 Time series analysis and Fourier transformation

(a) How could you test if there is a significant trend in X_t ? Explain a suitable test.

We can use these three tests mentioned below to find a significant trend:

Run test method:

1. Find the mean of the data set (μ)
2. Compare the values μ_i with μ (and assign + or -)
3. Count number of positive, negative and total runs
   - Total run (R) for negative (n_1) and positive (n_2)
4. Calculate μ and σ^2 (mean and variance)
   - $\mu = \frac{2 \cdot n_1 \cdot n_2}{n_1 + n_2} + 1$
   - $\sigma^2 = \frac{2 \cdot n_1 \cdot n_2 \cdot (2 \cdot n_1 \cdot n_2 - n_1 - n_2)}{(n_1 + n_1)^2 \cdot (n_1 + n_2 - 1)}$
5. Find Z = $\frac{R - \mu}{\sqrt{\sigma^2}}$
6. Find $Z_{1-\alpha/2}$
7. We have a trend if $Z > Z_{1-\alpha/2}$ where α is the significance level.

Mann-Kendall test method

1. State your null hypothesis
   - H_0 : no significant trend
2. State your null hypothesis
   - H_a : we have a significant trend
3. The Mann-Kendall statistic test is:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} = \text{sign}(x_j - x_i)$$

   - where n is the length of the data set
4. Before running the Mann-Kendall test we need to ensure that our data is
   - not collected seasonally (only summer og only winter months)
   - does not have any covariates
   - One data point per time period
5. The test calculates the difference in signs at every value to the preceeding value in the time series.
   - If we have a significant trend, the sign values will tend to constantly increase ($\theta > 0$).
   - Or opposite: negative trend when the sign values decrease constantly ($\theta < 0$).
   - Or no trend ($\theta = 0$)

$$\text{sign}(\theta) = \begin{cases} 1, & \theta > 0 \\ 0, & \theta = 0 \\ -1, & \theta < 0 \end{cases}$$

Linear-regression method

1. Fitting a linear regression equation with the independent variable (time T) and the dependent variable (hydrological variable)
2. State our null and alternative hypothesis
   - $H_0 = \beta \neq 0$
   - $H_a = \beta = 0$

3. We can use t-test to check if the slope is β different from zero or not (but we need also to test if the dataset is normality by using the Kolmogorov-Smirnov test). * **Our test statistic:** $t = \frac{\beta - 0}{s_\beta}$

    - **Rejection criteria:** H_0 is rejected if $|t| \geq t_{1-\alpha/2, n-2}$ where α is the significance level.

(b) The following three graphs show the absolute values for Fourier coefficients, defined as:

Figure a) shows the absolute values for the Fourier coefficients, since figure C has no noise at all, while figure b) shows a more frequent periodicity.