In [14]:
```
import numpy as np
import scipy.stats as st
```

1 Random variable parameter estimation

1) The expected value

$$E[X] = \sum_{i=1}^{n} X_i p_i = (-1) \cdot \frac{1}{3} + 3 \cdot \frac{1}{2} + 4 \cdot \frac{1}{6} = \frac{-2+9+4}{6} = \frac{11}{6} \approx 1.83 \quad (1)$$

2) The variance

$$Var(X) = E[X^2] - E[X]^2$$

$$E[X^2] = \sum_{i=1}^{n} X_i^2 p_i = (-1)^2 \cdot \frac{1}{3} + 3^2 \cdot \frac{1}{2} + 4^2 \cdot \frac{1}{6} = \frac{2+27+16}{6} = \frac{45}{6} = 7.$$

$$Var(x) = E[X^2] - E[X]^2 = 7.5 - 1.83^2 \approx 4.14$$

3) The mode

The mode is a measure of the value with the highest probability. In this case, the maximum $prob(X_i)$ is $\frac{1}{2}$ and corresponds to $X = 3$. The mode is therefore $X = 3$.

4) The coefficient of variation

The coefficient of variation is given by

$$CV = \frac{\sqrt{Var(X)}}{E[X]} \approx \frac{\sqrt{4.14}}{1.83} \approx 1.11 \quad (5)$$

2 Frequency analysis and linear regression

a)

We want to find the probability of observing at least one 100-years flood or larger within a period of of 10 years. A 100-years flood is defined in such a way that the probability of a flood of such magnitude (or larger) is $p = 1/100$ each year. The probability of *no* 100-years floods occuring during a 10 year period is therefore

$$p(\text{no 100-years floods}) = (1 - 1/100)^{10} \approx 0.904 \tag{6}$$

As our sample space consist of only either none 100-years floods occuring or at least one 100-years flood occuring, the probability of *at least one* 100-year flood occurring will therefore be

$$p(\text{at least one 100-years flood}) = 1 - p(\text{no 100-years floods}) \approx 1 - 0.904 = 0.$$

The probability of observing at least one 100-years flood or larger within a period of of 10 years is approximately 9.6 %.

b)

The assumptions behind a linear regression are:

- Linearity
- Normality
- Homoscedasity
- Independence

In figure 1A we see the simple linear regression between average runoff and median annual flood. We can therefore use this plot to evaluate whether linearity between the variables is a good assumption or not. We can see that a lot of our values are located the bottom left corner of the scatter plot (for small values), while there are a few more spread out values in the right half of the plot. There is some linearity to some extent, so the linearity assumption can only be said to be partly violated. However, the linear regression seems to be very influenced by some large outlier values. It could perhaps we more fruitful to consider the smaller values isolated.

In figure 1B (the Q-Q plot) we see to what extent our standardized residuals match the theoretical quantiles. We can therefore use this plot to evaluate normality. We want our standardized residuals to follow the theoretical quantiles in a 1-1 way. However, here we see that we do not get a straight diagonal line, we seem to violated the normality assumption here. An idea could be to consider the logarithm of our data and see if this is normally distributed, and use this for the linear regression instead.

To evaluate the homoscedasity and the independence assumptions, we would have to evaluate a scatter plot of the residuals and the autocorrelation respectively.

3 Confidence intervals

We have a sample of $N = 30$ random observations which produce a (sample) mean $\overline{x} = 145$ and (sample) variance $s^2 = 20$.

a)

We want to find the 95 % confidence interval on the mean assuming a normal distribution if

(i) *the true variance is unknown and estimated as 20*.

If the true variance is unknown, we need to take account of this uncertainty by applying the t-distribution instead of the normal distribution when finding the confidence interval, as the t-distribution has heavier tails than the normal distribution. The lower confidence limit l and the upper confidence limit u will therefore be given as

$$l = \overline{x} - t_{1-\alpha/2,N-1}s_{\overline{x}} \tag{8}$$
$$u = \overline{x} + t_{1-\alpha/2,N-1}s_{\overline{x}} \tag{9}$$

Here the t-distribution has $N - 1 = 29$ degrees of freedom, as we have a sample size of 30, and we use a significance level $\alpha = 0.05$ %, which corresponds to a confidence level of 95 %. We make use of the fact that the standard deviation of the mean $s_{\overline{x}}$ is given by $s_{\overline{x}} = \sqrt{s^2/N}$. Taking this into account, we can calculate the lower and upper confidence limits in the following way:

In [9]:
```
l = 145 - st.t.ppf(1-0.05/2,29)*np.sqrt(20/30)
u = 145 + st.t.ppf(1-0.05/2,29)*np.sqrt(20/30)
print("Lower confidence limit: ", l)
print("upper confidence limit: ", u)
```

```
Lower confidence limit:  143.33007698998662
upper confidence limit:  146.66992301001338
```

The 95 % confidence interval on the mean assuming with an unknown variance is therefore approximately [143.3 , 146.7]

(ii) *the true variance is 20*.

If the variance is known to be 20 we can use the regular normal distribution to find the confidence interval. This means that we can consider the sample variance s^2 as the true variance σ^2, and we can find the true variance of the mean as $\sigma^2_{\overline{x}} = \sigma^2/N$.

The lower confidence limit l and the upper confidence limit u will therefore be given as

$$l = \overline{x} - z_{1-\alpha/2}\sigma_{\overline{x}} \tag{10}$$
$$u = \overline{x} + z_{1-\alpha/2}\sigma_{\overline{x}} \tag{11}$$

We can calculate them in the following way:

In [10]:
```
l = 145 - st.norm.ppf(1-0.05/2)*np.sqrt(20/30)
u = 145 + st.norm.ppf(1-0.05/2)*np.sqrt(20/30)
print("Lower confidence limit: ", l)
print("upper confidence limit: ", u)
```

```
Lower confidence limit:  143.39969610788157
upper confidence limit:  146.60030389211843
```

The 95 % confidence interval on the mean assuming with an known variance is therefore approximately [143.4 , 146.6]

b)

We can see that the 95 % confidence interval with a known variance is slightly smaller than the one with an unknown variance. As stated before, this is due to the fact that we take more uncertainty into account by using the t-distribution, as this has heavier tails. However, it can be noted that the difference between the size of the confidence intervals as not very large. This tells us that a t-distribution with 29 degrees of freedom very closely resembles the normal distribution.

c)

We can find the 95 % confidence interval of the variance using a χ^2-distribution. The lower confidence limit l and the upper confidence limit u will then be given by

$$l = \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \tag{12}$$

$$u = \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \tag{13}$$

We can calculate them in the following way:

In [13]:
```
l = (30-1)*20/st.chi2.ppf(1 - 0.05/2,29)
u = (30-1)*20/st.chi2.ppf(0.05/2,29)
print("Lower confidence limit: ", l)
print("upper confidence limit: ", u)
```

```
Lower confidence limit:  12.685280051047778
upper confidence limit:  36.14366602272549
```

The 95 % confidence interval on the mean assuming with an known variance is therefore approximately [12.7 , 36.1]

4 Machine learning

a)

In machine learning, our goal is usually to make a model that "learns" (improves) from being fed more data, and which then can be generalized to be used for prediction or estimations based on new data. The degree by which the model performs when predicting data which has gone into the improvement of the model and the degree by which it performs when trying to make predicitons based on new data, are two very different measures. This is why we usually split our dataset into a training set and a test set.

The training set is the data that we use to improve our model. We feed it both the predictors and the values it should predict. The training error will be the error the model has when predicting the data from the training set. We can usually make this error smaller and smaller by increasing the complexity of our model and really "tuning it" to the data set we are training it on.

However, the training error will not be the same as the test error. The test error will be difference between what our model predict when being fed new data it has not been trained on, and the actual values it should have predicted. It is a measure of how well the model generalizes to new cases. To get a true estimate of this, it therefore very important that the data we use for testing the model has not gone into the "learning" of the model, but is independent. Only if we have split the data set into a training set and a test set from the very beginning, we are able to get good estimates of the test error, and not just the training error.

b)

How complex we want our model to be is connected to how well we want it to generalize. To be able to achieve a good generalisation, many machine learning algorithms have a parameter which controls the model complexity. As mentioned in exercise a), we can usually improve the *training error* of our model by increasing the complexity and finetuning our model to the test data. In the extreme case, this could lead to what we call overfitting. That would mean that our model is so finely tuned to the data that it performs really well on that specific dataset, but not so well on anything else. When looking at test error and training error together, the training error will usually always be higher than the training error. However, there comes a point when decreasing the training error of our model by increasing the complixity, happens on the expense of decreasing the test error, so that the test error will increase as the training error decreases. Controlling model complexity is a way of avoiding this.

5 Time series analysis and Fourier transformation

a)

We could use the linear regression method to test for a significant trend in the time series. This is a parametric test - we find a certain parameter for the trend using linear regression, and test if this parameter is significantly different from zero. In this method we assume that the residuals of our data are normally distributed. However, if we were unsure about if we wanted to make this assumption, we could have used a non-parametric test such as a Mann-Kendall test instead.

We proceed by performing a linear regression between our data points (t_i, X_{t_i}). This will gives some parameter a and b so that

$$\hat{X}_t = a + bt \tag{14}$$

We could then use a t-test to test the statistical signficance of the trend, in other words if our parameter b is significantly different from zero. This would be given by the test statistic

$$T = \frac{b - 0}{S_b},$$

where S_b is the standard deviation of b. Our null hypothesis would be that b is not significantly different from zero. We could use a two-tailed t-test, with a significance level α. We would have $n - 2$ degrees of freedom, as we have one model parameter. We will therefore detect a significant trend if we find that

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

as this is the rejection criteria for the null hypothesis. Here, $t_{1-\alpha/2,n-2}$ would be the critical value from the t-distribution.

We would first need to find the standard deviation of b. The standard deviation of b is given by

$$S_b = \frac{S}{\sqrt{\sum_i (t_i - \bar{t})^2}}$$

where S is the standard error of the regression

$$S = \sqrt{\frac{1}{n-2} \sum_i (X_{t_i} - \hat{X}_{t_i})^2}$$

We can see that the standard deviation of b depends on how large the time series that we have based the trend on is - larger time series, smaller standard deivation.

b)

In our signal, the clearest periodicity we see is a signal with a period of around $T = 5$ s. This corresponds to frequency of $f = 1/T = 0.2$ Hz. Since the signal is not completely smooth (with a single frequency), but appears to have some other frequencies as well, we expect there to be some noise in the frequency domain, but that the highest peak should appear around 0.2 Hz. We also expect this peak to be mirrored on the other side of the Nyquist frequency, which is 0.5 Hz in this case (half of the sampling freqency). (A) therefore appears to the the right plot of the Fourier transform of X_t.