1: Random variable parameter estimation

Have a random discrete variable X defined by:

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

Get a probability density function, f(x):

$$f(x) = \begin{cases} 1/3, & \text{if x} = -1 \\ 1/2, & \text{if x} = 3 \\ 1/6, & \text{if x} = 4 \end{cases}$$

a) Find the expected value:

Have the formula:

$$E(X) = \Sigma x * f(x)$$

By putting the values from the probability density function, we get that:

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

b) Find the variance:

Have the formula for variance, $V(X) = \sigma^2$:

$$\sigma^2 = \Sigma(x_i - \mu)^2 * p(x_i)$$

Where μ is the mean or expected value and $p(x_i)$ is the point probability of x_i.

$$\sigma^2 = \Sigma(x_i - \mu)^2 * p(x_i) = (-1 - \frac{11}{6})^2 * \frac{1}{3} + (3 - \frac{11}{6})^2 * \frac{1}{2} + (4 - \frac{11}{6})^2 * \frac{1}{6} = 4.14$$

```
In [14]: ((-1-11/6)**2)*(1/3)+((3-11/6)**2)*(1/2)+((4-11/6)**2*(1/6))
```

```
Out[14]: 4.138888888888889
```

c) Find the mode

The mode is 3. This is because 3 is the value which has the highest probability of occuring.

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?

The 100-years flood is the flood which has a probability of 0.01 of being observed one random year. Hence, the probability of a 100-year flood to not be observed one random year is $1 - 0.01 = 0.99$. Assuming that these events are independent, we get that the probability of a 100-year flood or larger to not be bserved in 10 years is 0.99^{10}. The probability that a 100-year flood or larger will be observed in 10 years is therefore $1 - 0.99^{10} = 0.096$.

```
In [20]: 1-0.99**10
```

```
Out[20]: 0.09561792499119559
```

3: Confidence intervals

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

i) the true variance is unknown and estimated as 20

Since the true variance is unknow, a t-value has to be used in the calculations. This t-value is found using a t-table. Have a variance,σ^2, of 20 and a mean, $\bar{x}$ of 145. The sample size, n, is 30. Get an estimated standard deviation,s_x, of $\sqrt{20} = 4.47$

Get that $s_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{4.47}{\sqrt{30}} = \frac{4.47}{\sqrt{30}} = 0.82$

Read from the t-table and get $t_{1-\alpha/2,n-1} = t_{0.975,30-1} = 2.045$

The lower, l, and upper, u, limit of the confidence interval is then given by:

$l = \bar{x} - t_{1-\alpha/2,n-1} * s_{\bar{x}} = 145 - 2.045 * 0.82 = 143.32$

$u = \bar{x} + t_{1-\alpha/2,n-1} * s_{\bar{x}} = 145 + 2.045 * 0.82 = 146.68$

ii) the true variance is 20

Since the true variance is known, a z-value must be used in the calculations.

Get that: $\sigma_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{4.47}{\sqrt{30}} = 0.82$

Read from z-table and get $z_{1-\alpha/2} = 1.96$

The lower and upper limit if the confidence interval is then given by:

$l = \bar{x} - z_{1-\alpha/2} * \sigma_{\bar{x}} = 145 - 1.96 = 143.04$

$u = \bar{x} + z_{1-\alpha/2} * \sigma_{\bar{x}} = 145 + 1.96 = 146.96$

b)

The reason for the different results in part (i) and (ii) is that when the true variance is known, there is a smaller uncertainty of getting the right values for x, i.e. a larger confidence interval. When the true variance is known, you have more information about the population and hence a smaller chance of getting a wrong value.

c) 95 % CI on the variance

Use the Chi-square distribution table.

$\chi^2_{\alpha/2,n-1} = \chi^2_{0.025,29} = 45.722$

$\chi^2_{1-\alpha/2,n-1} = \chi^2_{0.975,29} = 16.791$

$l = \frac{(n-1)s_x^2}{\chi^2_{\alpha/2,n-1}} = \frac{29*20}{45.722} = 12.69$

$u = \frac{(n-1)s_x^2}{\chi^2_{1-\alpha/2,n-1}} = \frac{29*20}{16.791} = 34.54$

The 95 % confidence interval for the variance is 12.69 to 34.54.

4: Machine learning

a)

In machine learning, it is common to split the dataset into a training set and a test set to get an impression of how exact your algorithm is. One can then use the training set to improve the algorithm and try to get a low training error, and then use the algorithm on the test set and compare it with the test error.

5: Time series analysis

a)

To test if there is a significant trend in X_t it is possible to use a linear regression and then use a t-test to see if the trend is significantly different from zero.