Exercise 13

Kolmogorov–Smirnov-Test

In this task, you investigate the similarity of the Poisson and Gaussian distributions using the Kolmogorov–Smirnov test.

a) What values do you have to choose for μ and σ of a Gaussian distribution so that it is as similar as possible to a Poisson distribution with expected value λ?

For $\mu = \lambda$ and $\sigma^2 = \lambda$ the gaussian distribution is as similiar as possible to a Poisson distribution.

b) Implement the two-sample Kolmogorov–Smirnov test for binned data.

In [1]:

```
import numpy as np

def KolSmi_test(data1, data2, alpha):
    n1, n2 = np.sum(data1), np.sum(data2)
    d = np.max(np.abs(data1/n1 - data2/n2)) #substracting the empirical distribution functions
    return np.sqrt((n1*n2)/(n1+n2))*d <= np.sqrt(np.log(2/alpha)/2) #checks wether the test is accepted or rejected
```

c) The two-sample Kolmogorov–Smirnov test checks the null hypothesis H_0, whether the two samples stem from the same probability distribution. Investigate at which expected value λ the Poisson and Gaussian distributions are so similar that the Kolmogorov–Smirnov test can no longer distinguish between the two. To do this, draw 10000 random numbers each from a Poisson distribution and from the corresponding Gaussian distribution for a λ to be tested. Consider the following:

• Round the values drawn from the Gaussian distribution to whole numbers.

• Use 100 bins each in the interval [$\mu - 5\sigma$, $\mu + 5\sigma$].

• Determine by iteration the value for λ from which you can no longer reject H_0 on the basis of the Kolmogorov–Smirnov test at a confidence level of $\alpha = 5\%$.

In [2]:

```
rng = np.random.default_rng(666)

import matplotlib.pyplot as plt

def test(lamda, alpha):

    data_p = rng.poisson(lam = lamda, size = 10000) #random numbers from a poisson distribution
    data_g = np.around(rng.normal(loc = lamda, scale = np.sqrt(lamda), size = 10000)) #rounded random numbers from
    #a normal distribution

    #bins for the poisson data
    bins1, limits, patches = plt.hist(data_p, bins = 100, range = (lamda-5*np.sqrt(lamda), lamda+5*np.sqrt(lamda)))
    #bins for the gaussian data
    bins2, limits, patches = plt.hist(data_g, bins = 100, range = (lamda-5*np.sqrt(lamda), lamda+5*np.sqrt(lamda)))
    plt.close()

    # test wether hypothesis is accepted
    return KolSmi_test(bins1, bins2, alpha)

l = np.linspace(1, 10, 100)

for i in range(len(l)):
    if test(l[i], 0.05) == True:
        print("Lambda_5.0 = ", l[i])
        break
```

```
Lambda_5.0 =  4.909090909090909
```

For a $\lambda \approx 5$ the null hyptothisis H_0, that the samples stem from the same probability distribution can no longer be rejected with a confidence level of $\alpha = 5\%$. This values seems to vary largely depending on the random numbers, that have been generated.

d) Determine the value for λ for the confidence levels α = 2.5 % and α = 0.1 % analogously.

In [3]:

```
for i in range(len(l)):
    if test(l[i], 0.025) == True:
        print("Lambda_2.5 = ", l[i])
        break
for i in range(len(l)):
    if test(l[i], 0.001) == True:
        print("Lambda_0.1 = ", l[i])
        break
```

```
Lambda_2.5 =  4.636363636363637
Lambda_0.1 =  3.090909090909091
```

For $\lambda \approx 4.64$ and a confidence level of $\alpha = 2,5\%$ and $\lambda \approx 3.09$ and a confidence level of $\alpha = 0,1\%$ the null hyptothesis can no longer be rejected. These values vary largly as well.

In [ ]: