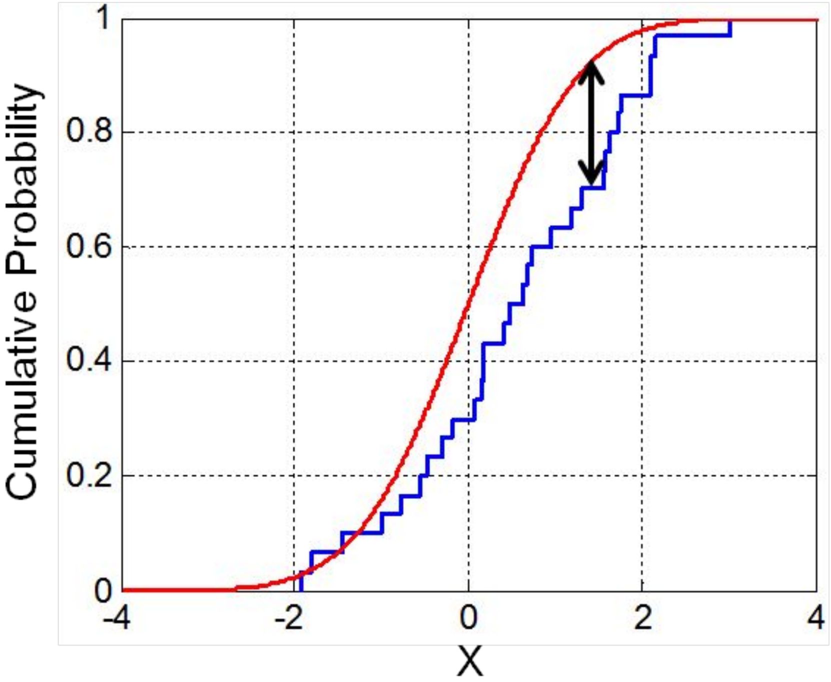
Kolmogorov-Smirnov test behaviour using Montecarlo Simulation

Dragos Tanasa - Statistical Learning

Kolmogorov-Smirnov test



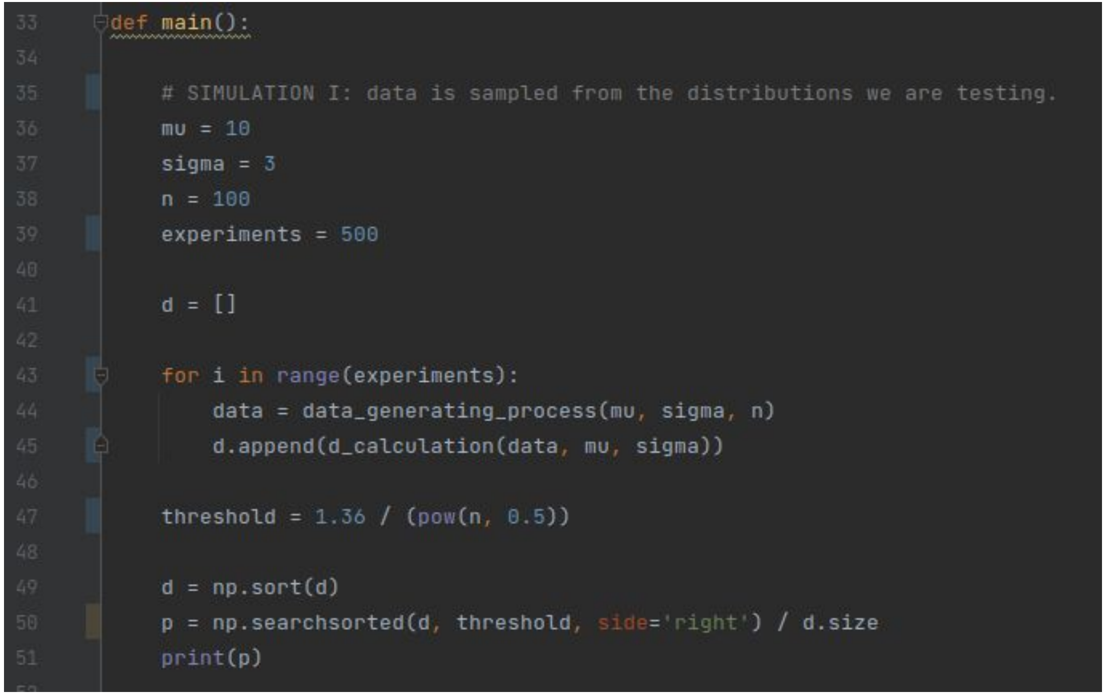
This is a statistical test that can be used to determine both if two sample came from the same distribution or if a sample is drawn from a particular probability distribution. In both cases this is done by confronting the cumulative distribution functions

$$F_n(x) = \frac{\text{number of (elements in the sample} \leq x)}{n} = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i)$$

The statistics of interest is D_n defined as:

$$D_n = \sup_x |F_n(x) - F(x)|$$

Code implementation



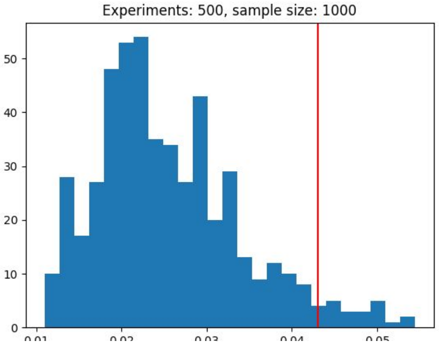
```python
def main():

    # SIMULATION I: data is sampled from the distributions we are testing.
    mu = 10
    sigma = 3
    n = 100
    experiments = 500

    d = []

    for i in range(experiments):
        data = data_generating_process(mu, sigma, n)
        d.append(d_calculation(data, mu, sigma))

    threshold = 1.36 / (pow(n, 0.5))

    d = np.sort(d)
    p = np.searchsorted(d, threshold, side='right') / d.size
    print(p)
```

For every iteration in the MM simulation a normal distributed sample is generated.

The d_calculation function outputs the value of D_n calculated between our sample and a Normal distribution with parameters given as an input to the function.

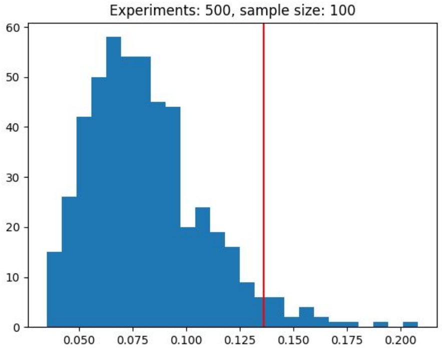
The threshold is set using tabulated values

Observations



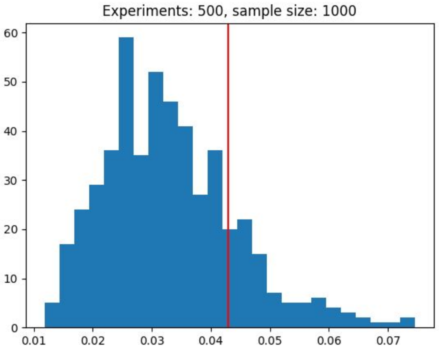
Case I:

Same distribution



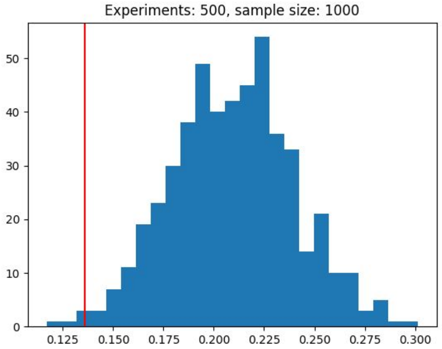
Case II:

Slightly different distribution, small sample size



Case III:

Slightly different distribution, bigger sample size



Case IV:

Different distributions

References:

- https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

- https://oak.ucc.nau.edu//rh83/Statistics/ks1/