1. Abstract: (200 words) What problem and major findings

This project aims at creating routines that fit an input of random variables to a set of continuous / discrete distributions to determine the best fitting distribution for the sample data. Input sample is tested against discrete distributions like Bernoulli, Poisson, Geometric and continuous distributions like Normal, Gamma and Weibull. We generate maximum likelihood estimates of the distributions that are very close to the actual parameters of the sample data set, and perform Chi-Squared goodness of fit tests on the sample data against every distribution to validate the goodness of fit of the routine. Additionally, we validate negative scenarios, for e.g., an input geometric data should fail the goodness of fit test for a normal distribution.

2. Organization of the report:

The rest of the report is organized as follows:
The main findings and logic for maximum likelihood estimations and goodness of fit are elaborated in Chapter 3. Discussions on testing scenarios to validate estimations and Chi-squared goodness of fit are discussed in Chapter 4. We conclude in Chapter 5 with future work.

3. Main Findings

Distributions under consideration:

- Discrete: Poisson, Geometric, Bernoulli
- Continuous: Normal, Gamma, Weibull

The above discrete distributions and continuous distributions were tried to fit a Maximum Likelihood Estimate for. Once the MLE estimate gives the parameters of the best distribution, we evaluate the fit using the Goodness of fit (Chi-squared) test.

Maximum Likelihood estimate: MLE is a technique used to estimate the parameters of a distribution. In this technique, we find the parameter values that maximize the likelihood of making the observations given the parameters. Specifically, we differentiate the distribution wrt to the parameters and equate it to zero to find the best parameters that satisfy the optimal fit.

Goodness of fit test: We use the chi square test to assess the goodness of fit for the assumed distribution. For the example of a normal distribution hypothesis testing, the null hypothesis is that the distribution is normal and the alternate hypothesis is that we have sufficient evidence to prove that the distribution is non-normal. Therefore, a p value of less than 0.05 or a Test statistic greater than Critical statistic (as shown in the table below) indicates rejection of the null hypothesis for the distribution. The logic for the test is coded by equally sampling input data into bins and calculating the expected (Ei)/ observed frequencies(Oi) of data in each range. The Chi-squared statistic is calculated using the below logic:

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i$$

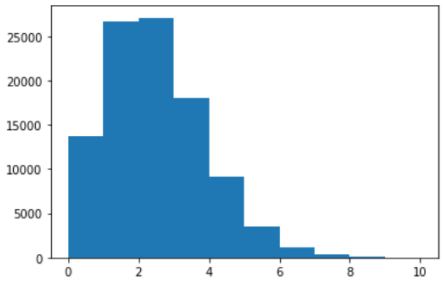
I. Experiment 1: Goodness of fit for sample data generated using numpy.random.distributions

Basic plot of the numbers generated-

Results:

Discrete Distributions:

Poisson (2):



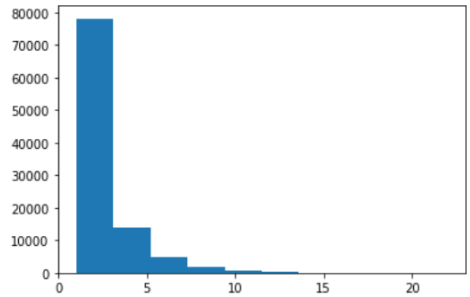
Output from routine:

	Distribution	Maximum Likelihood Estimators	Test Statistic	Critical Statistic	PValue
2	Poisson	[2.0042500000000003]	8.584892e+00	19.675138	0.660148
0	Bernoulli	[2.00425]	1.336279e+08	100734.728821	0.000000
1	Exponential	[0.49893975302482224]	1.054535e+05	100734.728821	0.000000
3	Geometric	[0.4989397530248223]	8.978607e+04	19.675138	0.000000
4	Normal	[2.0042500000000003, 2.0065919375]	inf	9.487729	0.000000
5	Gamma	[2.0019107958266673, 1.00116848571785]	inf	14.067140	0.000000
6	Weibull	[-2.557083549241807, 0.0]	7.000000e+05	11.070498	0.000000

The MLE estimator is very close to the actual parameters (rate=2) used to generate the input data.

From the p value as well as the Test statistic in relation to the Critical statistic, we see here that the randomly generated Poisson data fails to reject the null hypothesis that the distribution is Poisson. All other distributions reject the hypothesis that it belongs to other distributions.

Geometric (0.4):



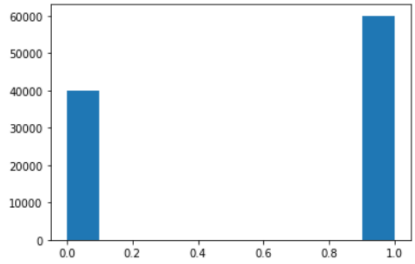
Output from routine:

	Distribution	Maximum Likelihood Estimators	Test Statistic	Critical Statistic	PValue
3	Geometric	[0.39916972696790676]	2.300587e+01	33.924438	0.401398
0	Bernoulli	[2.5052]	1.153330e+10	100734.728821	0.000000
1	Exponential	[0.39916972696790676]	1.537380e+05	100734.728821	0.000000
2	Poisson	[2.5052000000000003]	9.668135e+09	33.924438	0.000000
4	Normal	[2.5052000000000003, 3.79777296]	inf	9.487729	0.000000
5	Gamma	[1.6525545645045618, 1.5159559955292987]	inf	14.067140	0.000000
6	Weibull	[1.4496982298866146, 2.7928556823180886]	6.999720e+05	11.070498	0.000000

The MLE estimator is very close to the actual parameters (Prob=0.4) used to generate the input data.

From the p value as well as the Test statistic in relation to the Critical statistic, we see here that the randomly generated geometric data fails to reject the null hypothesis that the distribution is geometric. All other distributions reject the hypothesis that it belongs to other distributions.

Bernoulli (0.6):



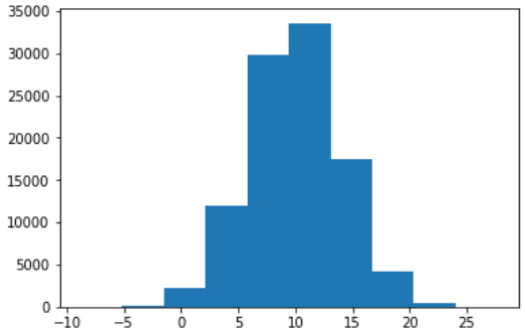
Output from routine:

	Distribution	Maximum Likelihood Estimators	Test Statistic	Critical Statistic	PValue
0	Bernoulli	[0.59938]	0.000000e+00	100734.728821	1.0
1	Exponential	[1.6683906703593714]	4.097528e+05	100734.728821	0.0
2	Poisson	[0.59938]	2.620296e+04	3.841459	0.0
3	Normal	[0.59938, 0.24012361560000003]	inf	9.487729	0.0
4	Gamma	[1.49613099695472, 0.40062]	inf	14.067140	0.0
5	Weibull	[-32.056424851445406, 0.0]	7.000000e+05	11.070498	0.0

The MLE estimator is very close to the actual parameters (Prob = 0.6) used to generate the input data.
From the p value as well as the Test statistic in relation to the Critical statistic, we see here that the randomly generated Bernoulli data fails to reject the null hypothesis that the distribution is Bernoulli. All other distributions reject the hypothesis that it belongs to other distributions.

Continuous Distributions:

Normal (10,16):



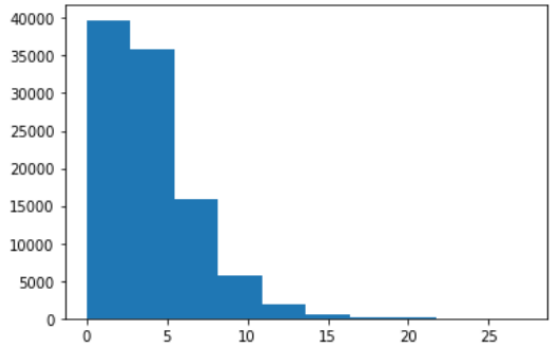
Output from routine:

	Distribution	Maximum Likelihood Estimators	Test Statistic	Critical Statistic	PValue
4	Normal	[10.000757274862552, 16.051779409940618]	3.488275e+00	9.487729	0.625163
0	Bernoulli	[10.000757274862508]	6.659213e+08	100734.728821	0.000000
1	Exponential	[0.09999242782479671]	8.518087e+04	100734.728821	0.000000
2	Poisson	[10.00075727486251]	7.797899e+08	100735.732499	0.000000
3	Geometric	[0.09999242782479627]	4.444475e+08	100735.732499	0.000000
5	Gamma	[6.230782489371767, 1.605056394108038]	4.137097e+03	14.067140	0.000000
6	Weibull	[2.4961410580227272, nan]	1.000000e+05	11.070498	0.000000

The MLE estimators are very close to the actual parameters (Mean = 10, Variance = 16) used to generate the input data.

From the p value as well as the Test statistic in relation to the Critical statistic, we see here that the randomly generated normal data fails to reject the null hypothesis that the distribution is normal. All other distributions reject the hypothesis that it belongs to other distributions.

Gamma (2,2):

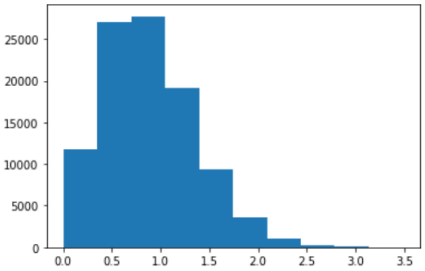


Output from routine:

	Distribution	Maximum Likelihood Estimators	Test Statistic	Critical Statistic	PValue
5	Gamma	[2.004479937744833, 1.9983688996709732]	3.793689e+00	14.067140	0.924459
0	Bernoulli	[4.005690367603692]	1.953827e+09	100734.728821	0.000000
1	Exponential	[0.24964485724797195]	1.539530e+04	100734.728821	0.000000
2	Poisson	[4.005690367603692]	1.173871e+09	100735.732499	0.000000
3	Geometric	[0.24964485724797253]	1.552398e+09	100735.732499	0.000000
4	Normal	[4.005690367603684, 8.004847052330788]	1.236673e+04	9.487729	0.000000
6	Weibull	[1.4823509417478555, 4.443895715253564]	6.752521e+05	11.070498	0.000000

The MLE estimator is very close to the actual parameters (Shape=2, Scale=2) used to generate the input data.
From the p value as well as the Test statistic in relation to the Critical statistic, we see here that the randomly generated gamma data fails to reject the null hypothesis that the distribution is gamma. All other distributions reject the hypothesis that it belongs to other distributions.

Weibull (2,1):



Output from routine:

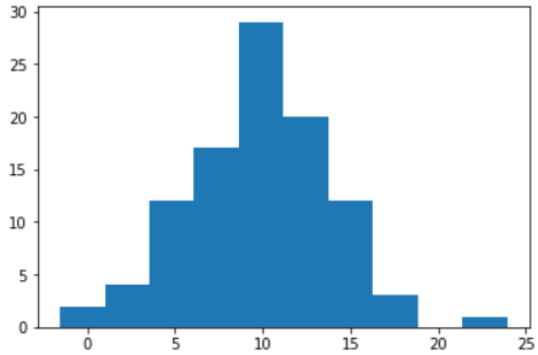
	Distribution	Maximum Likelihood Estimators	Test Statistic	Critical Statistic	PValue
5	Weibull	[1.996247436788764, 0.9986327032543378]	7.110640e+00	11.070498	0.417452
0	Bernoulli	[0.8850592859921679]	5.119640e+09	100734.728821	0.000000
1	Exponential	[1.1298678131815558]	3.801459e+04	100734.728821	0.000000
2	Poisson	[0.885059285992168]	3.915448e+09	100735.732499	0.000000
3	Normal	[0.8850592859921725, 0.21473065106389108]	3.147048e+03	9.487729	0.000000
4	Gamma	[3.6479651872703625, 0.24261725114060909]	1.863619e+03	14.067140	0.000000

The MLE estimator is very close to the actual parameters (Shape=2, Scale=1) used to generate the input data.
From the p value as well as the Test statistic in relation to the Critical statistic, we see here that the randomly generated Weibull data fails to reject the null hypothesis that the distribution is Weibull. All other distributions reject the hypothesis that it belongs to other distributions.

II. Experiment 2: Effect of changing the number of observations

The above results were for 100000 random samples, when the same experiment is repeated with just 100 data points, we get the following output:

Normal (10,16):



Output from routine:

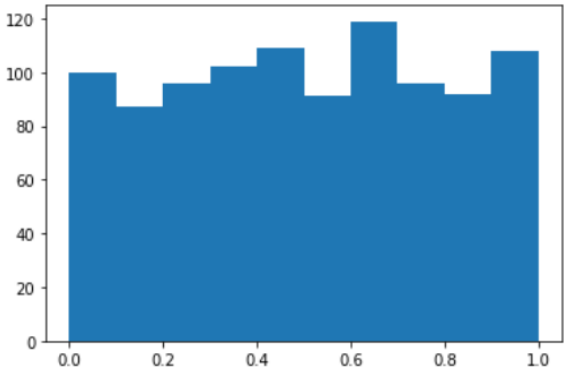
	Distribution	Maximum Likelihood Estimators	Test Statistic	Critical Statistic	PValue
3	Normal	[9.869560180725536, 17.010834663513748]	6.169578	9.487729	2.900663e-01
4	Gamma	[5.726245659767201, 1.7235656252174798]	19.199702	14.067140	2.354741e-02
5	Weibull	[2.380962663119591, nan]	100.000000	11.070498	1.078798e-18
1	Exponential	[0.10132163760984203]	112.100000	122.107735	1.408688e-20
2	Poisson	[9.86956018072554]	799.604655	123.225221	4.738588e-110
0	Bernoulli	[9.86956018072554]	68441.335504	122.107735	0.000000e+00

We see that the decision of distribution is not very well-defined owing to lack of clarity due to the low number of points. Even though the result is not well defined, the highest p-value is for Normal distribution, which considers this data a close fit to Normal. The variance in MLE estimators is also higher than the estimators with large number of samples.

III. Experiment 3: Fitting a Uniform Distribution

All the above experiments were done using input distributions for which routines were defined. Now, we are trying to fit a uniformly distributed data through our routines. Since we have not defined a routine that will identify uniform distribution, we are interested in observing these results.

Uniform (0,1):



Output from routine:

	Distribution	Maximum Likelihood Estimators	Test Statistic	Critical Statistic	PValue
3	Normal	[0.5070268435589639, 0.08259977442337203]	54.496378	9.487729	1.656965e-10
1	Exponential	[1.972282163564988]	310.640000	1072.605834	2.235255e-62
4	Gamma	[3.1123114062237085, 0.1629100618097081]	597.285548	14.067140	8.009061e-123
0	Bernoulli	[0.5070268435589639]	492973.156441	1072.605834	0.000000e+00
2	Poisson	[0.5070268435589639]	602283.602831	1073.642651	0.000000e+00
5	Weibull	[1.6371178367970114, 0.5590533174179465]	2160.272000	11.070498	0.000000e+00

We see that goodness of fit test has very low probability for this input distribution. Our goodness of fit tests works well since we are not expecting any of the defined distributions to match with a uniform distribution.

4. Discussions

We have successfully implemented MLE for estimating hyperparameters and assessed the resultant goodness of fit using chi square test.
We also saw the effect of having a lesser number of datapoints on the conclusion of the hypothesis testing. Lower number of datapoints cause the decision to be ambiguous.

5. Conclusions and Future Work:

We infer that maximum likelihood estimators give the best estimates for input data distributions. Also, Chi-squared goodness of fit tests work well when data has large number of samples. In future, we should explore other goodness of fit tests on more distributions. This will help us utilize specific tests based on the nature of data samples to achieve best outcomes.