HW4 - ISYE 6644

Marcos Grillo

March 2023

1

1.1

For a log-normal distribution:

Since the log-data is IID normal, this means that, for a sample x_i of data, with mean $\bar{X}$ and variance S^2, if we were to take the log of these values, $ln(x_i)$ would result in a normal distribution, with mean $\bar{Y}$ and variance s^2:

$$\mu \approx \bar{Y}$$
$$\sigma^2 \approx s^2$$

Therefore, the estimated parameters here are:

$$\hat{\mu} = \bar{Y}$$
$$\hat{\sigma^2} = s^2$$

The MLE's for a log-normal distribution are:

$$\begin{aligned}\hat{\mu} &= \frac{\sum_{i=1}^{n} lnx_i}{n} \\ &= \bar{Y}\end{aligned} \tag{1}$$

$$\begin{aligned}\hat{\sigma^2} &= \frac{\sum_{i=1}^{n}(lnx_i - \hat{\mu})^2}{n} \\ &= s^2\end{aligned} \tag{2}$$

Therefore, for a log-normal distribution, the MLE's are equivalent to the estimators obtained from the method of moments as n goes to infinity.

1.2

The mean and variance of a binomial distribution, with sample x_i of size n and probability of success p, are:

$$E(X) = np$$

$$Var(X) = np(1-p)$$

The same mean and variance are:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

Using the method of moments:

$$np \approx \bar{X}$$

$$p \approx \frac{\bar{X}}{n}$$

The MLE's for a binomial distribution are:

$$\hat{p} = \frac{x}{n}$$

In this case, the estimators derived from the method of moments and the MLE's are equal.

2

The likelihood function, for $x_i \geq y$:

$$L(y, \beta) = \prod_{i=1}^{n}(\frac{1}{\beta})e^{-\frac{(x_i - y)}{\beta}}$$

Taking the log to get the log-likelihood:

$$l(y, \beta) = \sum_{i=1}^{n}(-ln(\beta) - \frac{x_i - y}{\beta})$$

Calculating MLE's, starting with y, we recall that $x_i \geq y$. Therefore, the MLE of y is when y is equal to the smallest of the x_i observations, or:

$$\hat{y} = min(x_1, x_2, ...x_n)$$

Now, for β:

$$\begin{aligned}\frac{d}{d\beta}l(y,\beta) &= \frac{d}{d\beta}\sum_{i=1}^{n}(-ln(\beta) - \frac{x_i - y}{\beta}) \\ &= \sum_{i=1}^{n}\frac{-1}{\beta} - \frac{-x_i + y}{\beta^2}\end{aligned} \tag{3}$$

Setting this equal to 0:

$$0 = \sum_{i=1}^{n}\frac{-1}{\beta} - \frac{-x_i + y}{\beta^2}$$

$$\sum_{i=1}^{n}\frac{-x_i + y}{\beta^2} = \sum_{i=1}^{n}\frac{-1}{\beta}$$

$$\sum_{i=1}^{n}\frac{-x_i + y}{\beta^2} = \frac{-n}{\beta}$$

$$\frac{1}{n}(x_i + y) = \frac{-n(\beta^2)}{\beta}$$

$$\frac{1}{n}(x_i + y) = -n(\beta)$$

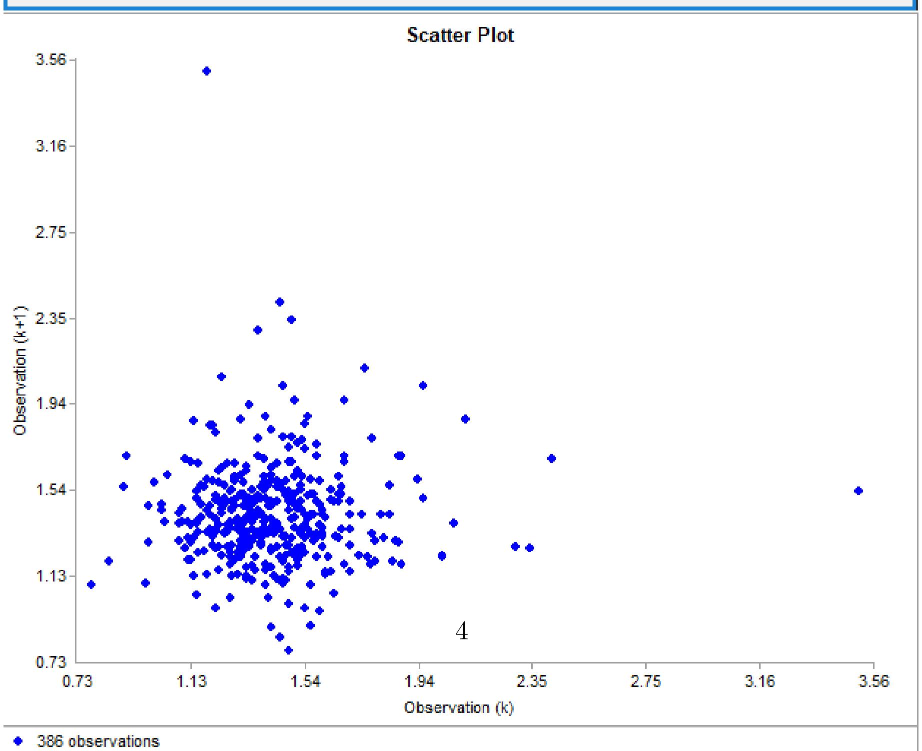
$$\frac{\frac{1}{n}(x_i + y)}{n} = \hat{\beta}$$

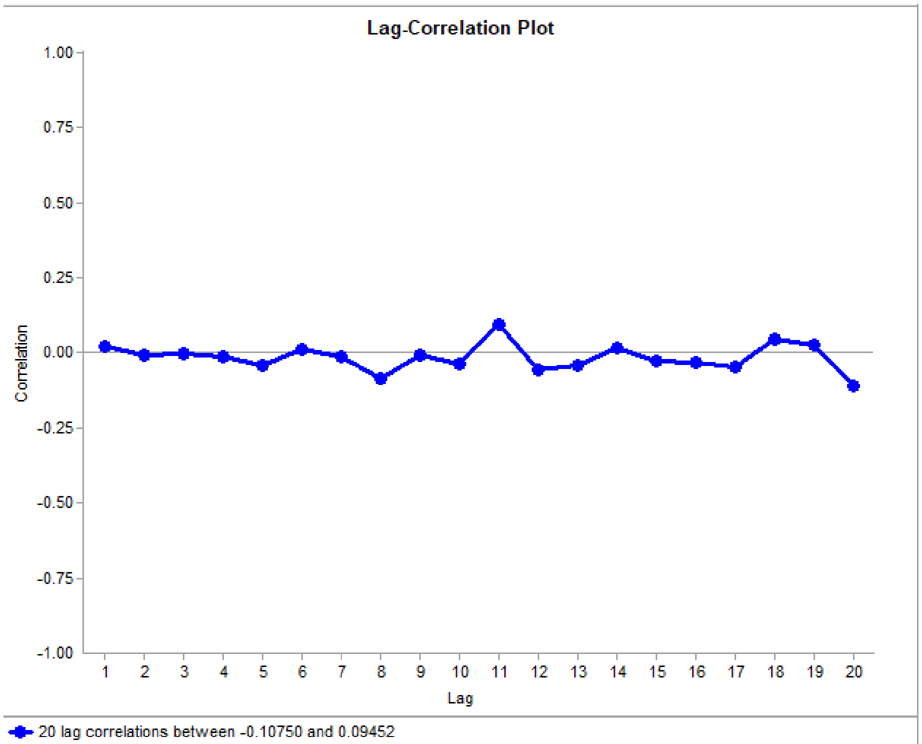
$$x_i + y = \hat{\beta}$$

3

Data-Summary Table

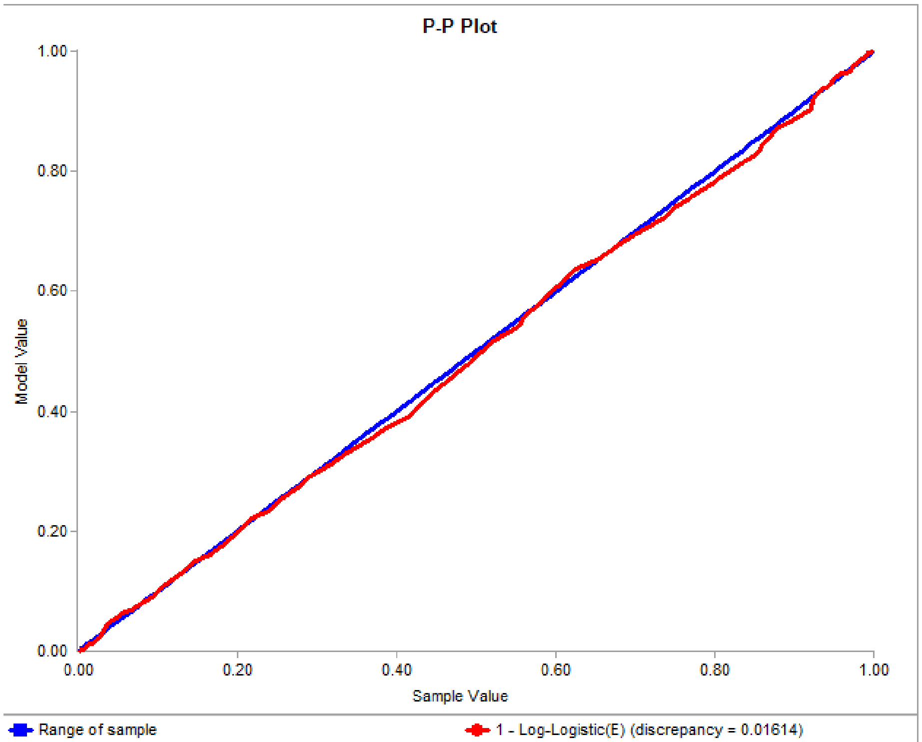
Data Characteristic	Value
Source file	6644hw4s23-A
Observation type	Real valued
Number of observations	386
Minimum observation	0.78000
Maximum observation	3.51000
Mean	1.42448
90.0% c.i. half-length	0.02052
Median	1.40000
Variance	0.05980
Coefficient of variation	0.17167
Skewness	2.15606
Kurtosis	14.41447
1st percentile	0.90850
5th percentile	1.11417
10th percentile	1.16833
90th percentile	1.67375
95th percentile	1.82500
99th percentile	2.11600





Relative Evaluation of Candidate Models

Model	Relative Score	Parameters	
1 - Log-Logistic(E)	99.14	Location	0.26079
		Scale	1.14009
		Shape	9.58568
2 - Log-Logistic	97.41	Location	0.00000
		Scale	1.40234
		Shape	11.79421
3 - Pearson Type V	85.34	Location	0.00000
		Scale	56.03732
		Shape	40.35672



Anderson-Darling Test with Model 1 - Log-Logistic(E)

Sample size	386
Test statistic	0.25175

Note: No critical values exist for this special case. The following critical values are for the case where all parameters are known, and are conservative.

	Critical Values for Level of Significance (alpha)					
Sample Size	0.250	0.100	0.050	0.025	0.010	0.005
386	1.248	1.933	2.492	3.070	3.857	4.500
Reject?	No					

Kolmogorov-Smirnov Test with Model 1 - Log-Logistic(E)

Sample size	386
Normal test statistic	0.02852
Modified test statistic	0.56034

Note: No critical values exist for this special case. The following critical values are for the case where all parameters are known, and are conservative.

	Critical Values for Level of Significance (alpha)				
Sample Size	0.150	0.100	0.050	0.025	0.010
386	1.131	1.216	1.349	1.471	1.618
Reject?	No				

Simio expression for a Extended Log-Logistic expression:

Random.LogLogistic(9.58568, 1.14009) + 0.26079

As seen here, the data given is a good fit for an Extended Log-Logistic distribution, failing to reject in both the Anderson-Darling and the Kolmogorov-Smirnov tests, though both of these tests are conservative. The P-P plot also seems to align well with the data for this distribution.

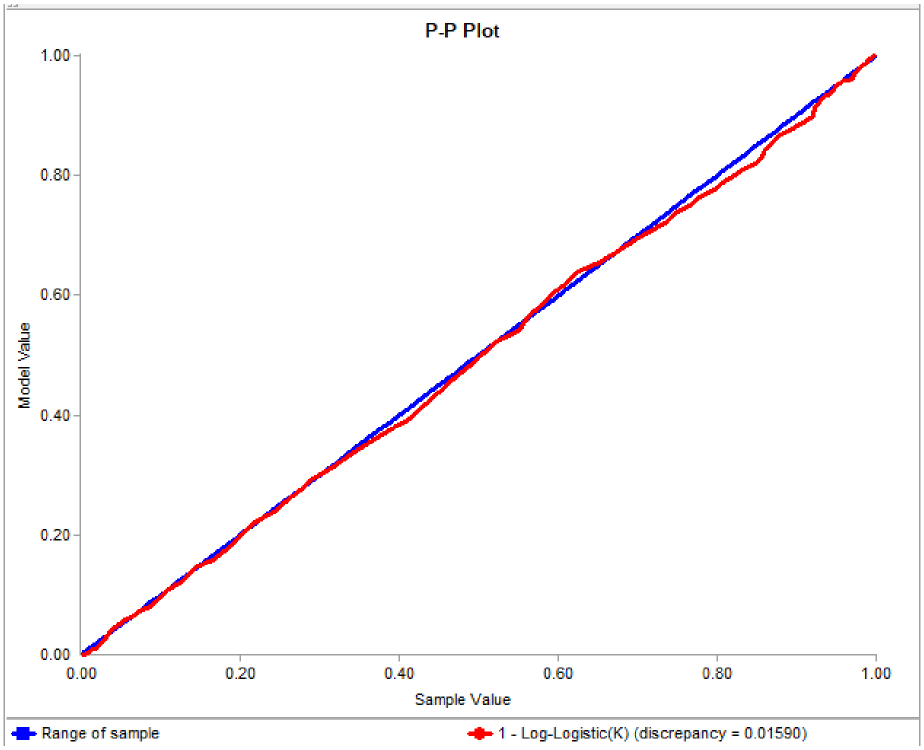
The density function in this case would be:

$$\begin{aligned} f(x;\alpha,\beta) &= \frac{(9.58568/1.14009)(x/1.14009)^{9.58568-1}}{(1+(x/1.14009)^{9.58568})^2} + 0.26079 \\ &= \frac{2.72786x^{8.58568}}{(0.284577x^{9.58568}+1)^2} + 0.26079 \end{aligned} \tag{4}$$

If setting a lower bound of 0.5, the following is seen:

Relative Evaluation of Candidate Models

Model	Relative Score	Parameters	
1 - Log-Logistic(K)	100.00	Location	0.50000
		Scale	0.89879
		Shape	7.52493
2 - Log-Laplace(K)	89.58	Location	0.50000
		Scale	0.90000
		Shape	5.39971
3 - Pearson Type VI(K)	88.54	Location	0.50000
		Scale	1.43118
		Shape #1	26.66944
		Shape #2	42.30567



Anderson-Darling Test with Model 1 - Log-Logistic(K)

Sample size 386
Test statistic 0.31655

Note: The following critical values are exact.

Sample Size	Critical Values for Level of Significance (alpha)					
	0.250	0.100	0.050	0.025	0.010	0.005
386	0.426	0.563	0.660	0.769	0.905	1.009
Reject?	No					

Kolmogorov-Smirnov Test with Model 1 - Log-Logistic(K)

Sample size 386
Normal test statistic 0.03028
Modified test statistic 0.59496

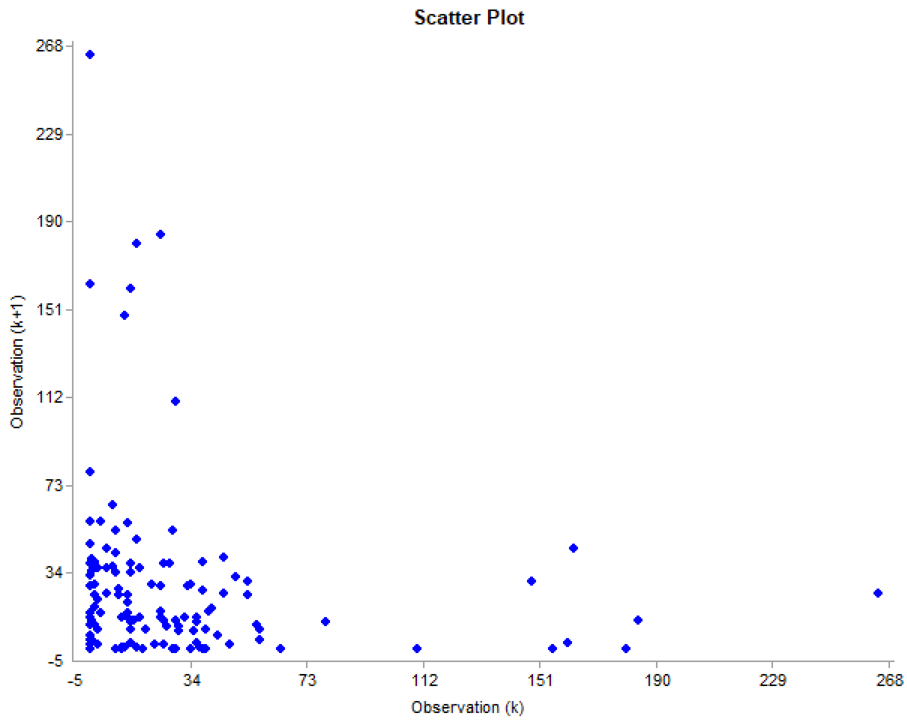
Note: The following critical values are exact.

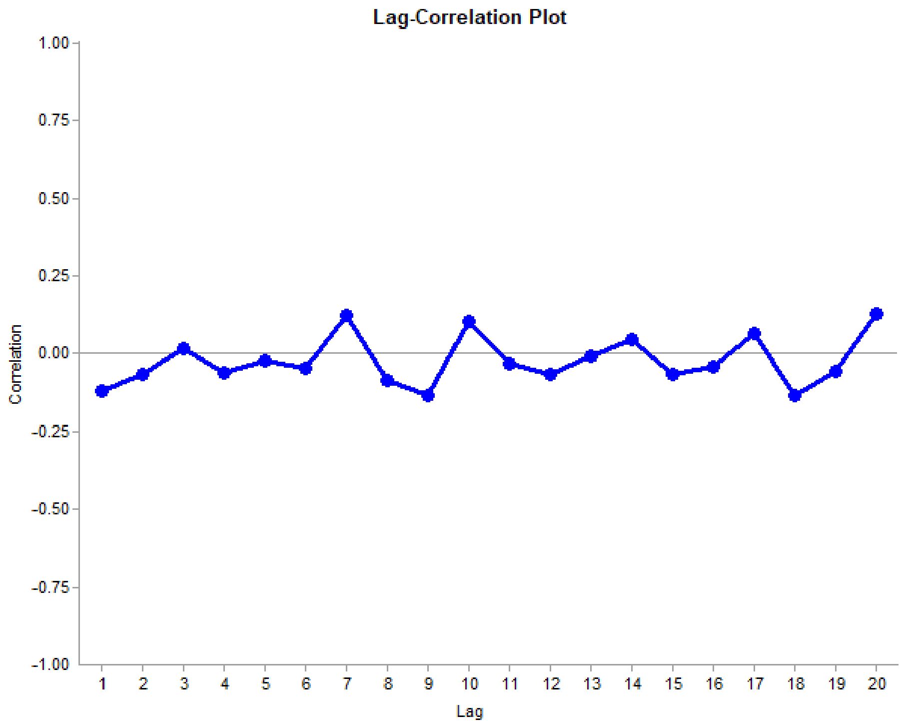
Sample Size	Critical Values for Level of Significance (alpha)			
	0.100	0.050	0.025	0.010
50	0.708	0.770	0.817	0.873
infinity	0.715	0.780	0.827	0.886
Reject?	No			

Given that the best fit here is a Kumaraswamy Log-Logistic distribution, both the Anderson-Darling and the Kolmogorov-Smirnov tests have non-conservative estimates, as well as a higher relative score as evaluated by ExpertFit, it would be appropiate to say that with a lower bound of 0.5 days, a better fit is seen.

4

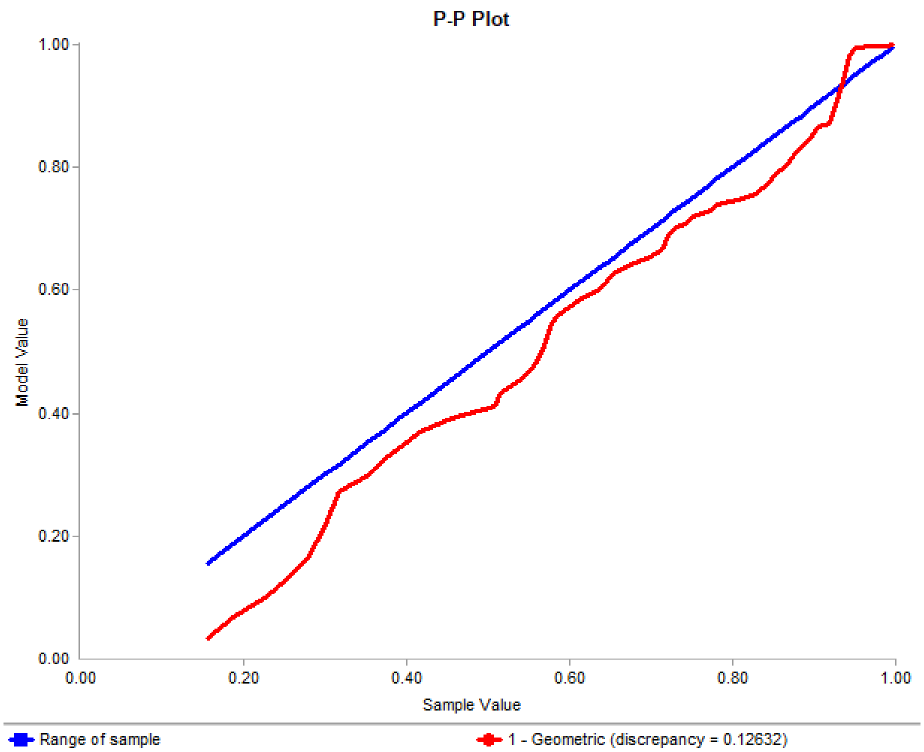
Data Characteristic	Value
Source file	6644hw4s23-B
Observation type	Integer valued
Number of observations	131
Minimum observation	0
Maximum observation	264
Mean	27.75573
90.0% c.i. half-length	6.00036
Median	14.00000
Variance	1,718.53987
Lexis ratio (var./mean)	61.91659
Skewness	3.18780
Kurtosis	11.74032





Relative Evaluation of Candidate Models

Model	Relative Score	Parameters	
1 - Geometric	83.33	Probability	0.03478
2 - Negative Binomial	83.33	Probability	0.03478
		Success	1
3 - Poisson	22.22	Lambda	27.75573



Equal-Width Chi-Square Test with Model 1 - Geometric

Upper endpoint of first interval	9
Interval width (values per interval)	10
Number of intervals (grouped/original)	27/27
Number of intervals with fewer than five expected observations per interval	21
Test statistic	175.25787

Degrees of Freedom	Observed Level of Significance	Critical Values for Level of Significance (alpha)				
		0.25	0.15	0.10	0.05	0.01
25	0.000	29.339	32.282	34.382	37.652	44.314
26	0.000	30.435	33.429	35.563	38.885	45.642
	Reject?	Yes				

Simio expression for a Geometric expression:

Random.Geometric(0.03478)

In this case, a geometric distribution doesn't seem to be a good fit, this is based on the fact that the Chi-Squared test rejects the null hypothesis that the data given fits a Geometric distribution. The P-P plot also seems to be a poor fit for a Geometric distribution, with significant discrepancy all along the value range, specially in the lower tails.

5

The following are the fitted parameters:

Parameter	Estimated Value	Standard Deviation
size	0.51956510122101	0.0634944866704382
mu	27.7557251985778	3.39565939739202

We can get p as follows:

$$\begin{aligned} p &= \frac{r}{\mu + r} \\ &= \frac{0.5196}{27.7557 + 0.5196} \\ &= 0.0184 \end{aligned} \tag{5}$$

And we can get the variance by:

$$\begin{aligned} \sigma^2 &= r(\frac{1-p}{p^2}) \\ &= 0.5196(\frac{1-0.0184}{0.0184^2}) \\ &= 1506.4962 \end{aligned} \tag{6}$$

The fitted parameters here are not similar to the sample moments. ExpertFit suggests a p value of 0.03478 (notably, both for its fit to a Geometric and a Negative Binomial distribution), which differs significantly from the p value given here of 0.0184.

6

6.1

Expertfit was not able to find any reasonably fitting distributions for this dataset. This is shown by the dataset failing to maintain the null hypothesis in both the Anderson-Darling and the Kolmogorov-Smirnov tests for the Beta distribution, suggested by ExpertFit:

Anderson-Darling Test with Model 1 - Beta

Sample size	2,590
Test statistic	8.78285

Note: No critical values exist for this special case. The following critical values are for the case where all parameters are known, and are conservative.

Sample Size	Critical Values for Level of Significance (alpha)					
	0.250	0.100	0.050	0.025	0.010	0.005
2,590	1.248	1.933	2.492	3.070	3.857	4.500
Reject?	Yes					

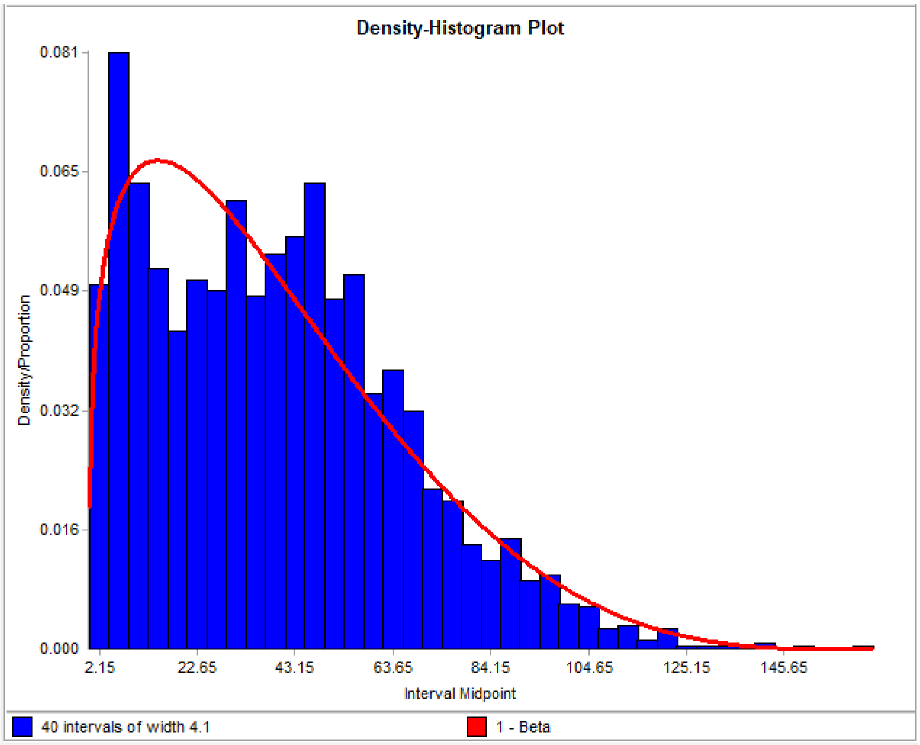
Kolmogorov-Smirnov Test with Model 1 - Beta

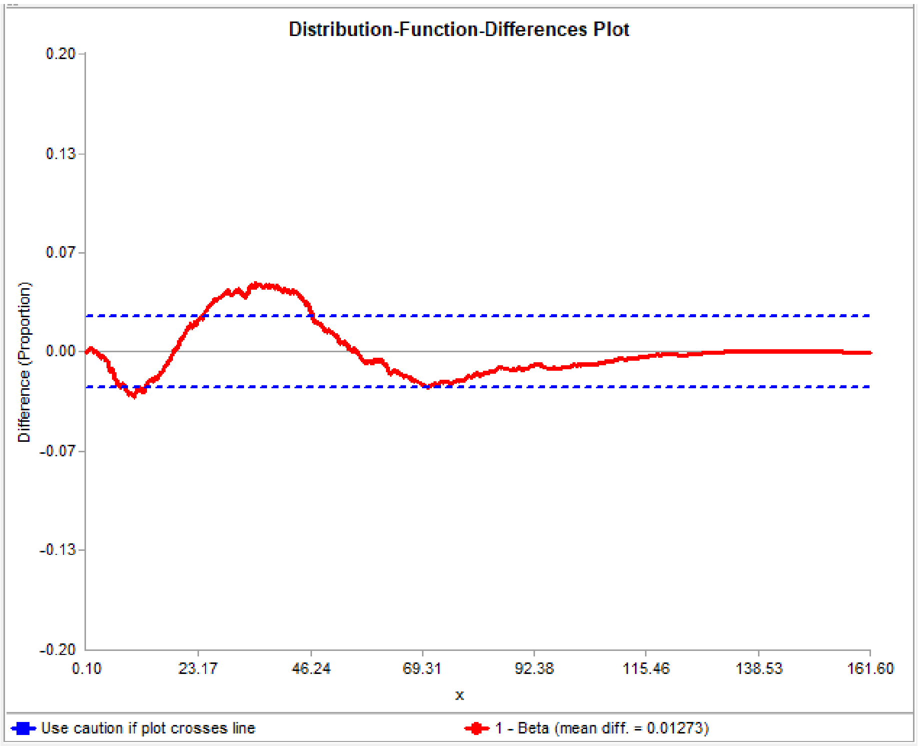
Sample size	2,590
Normal test statistic	0.04720
Modified test statistic	2.40201

Note: No critical values exist for this special case. The following critical values are for the case where all parameters are known, and are conservative.

Sample Size	Critical Values for Level of Significance (alpha)				
	0.150	0.100	0.050	0.025	0.010
2,590	1.135	1.221	1.355	1.476	1.624
Reject?	Yes				

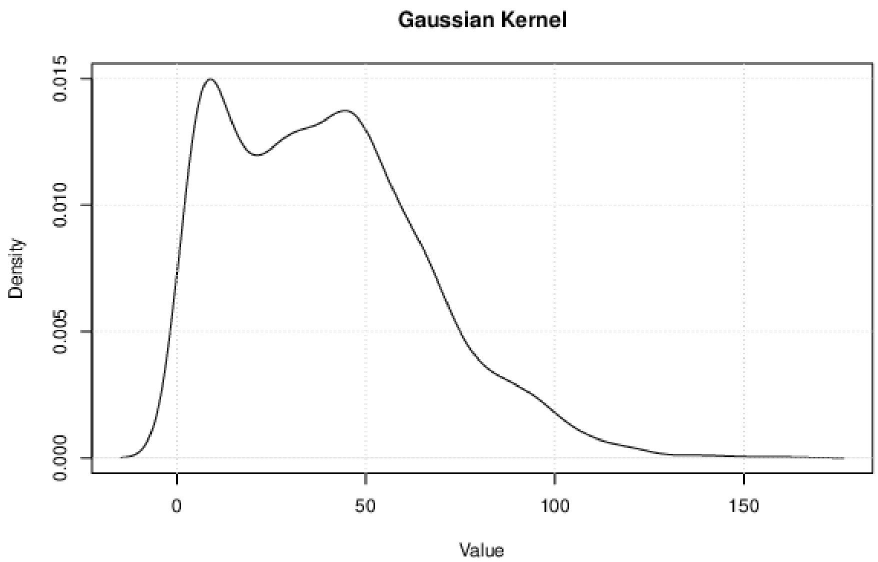
One potential explanation for this poor fit may be due to the poor fit around the 20's range. This can be shown by the density-histogram and Distribution-Function-Differences plots shown below.


Density-Histogram Plot
Density/Proportion
0.081
0.065
0.049
0.032
0.016
0.000
2.15
22.65
43.15
63.65
84.15
104.65
125.15
145.65
Interval Midpoint
40 intervals of width 4.1
1 - Beta


Distribution-Function-Differences Plot
0.20
0.13
0.07
0.00
-0.07
-0.13
-0.20
Difference (Proportion)
0.10
23.17
46.24
69.31
92.38
115.46
138.53
161.60
x
Use caution if plot crosses line
1 - Beta (mean diff. = 0.01273)

6.2

The KDE using a Gaussian kernel is shown below:

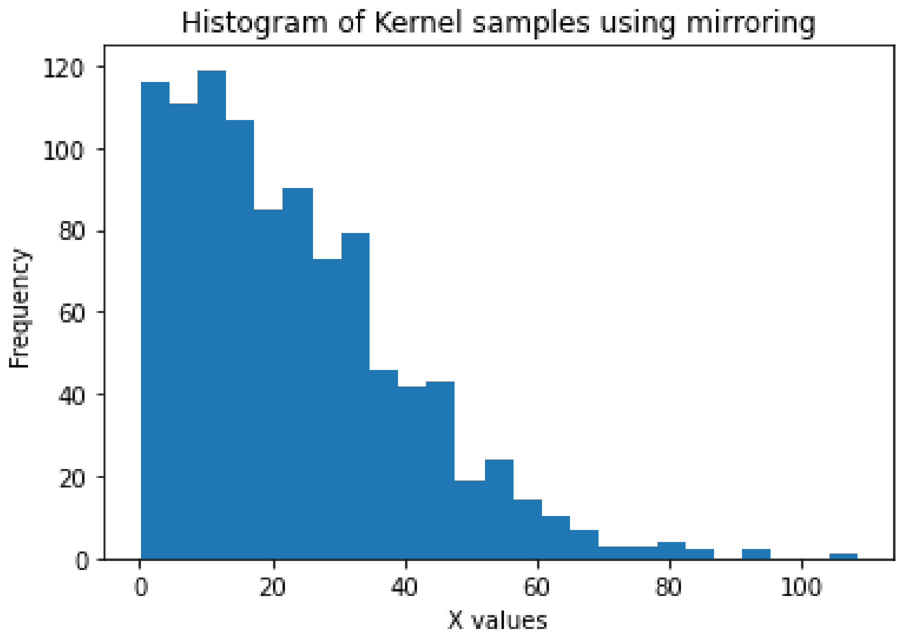


With the following values:

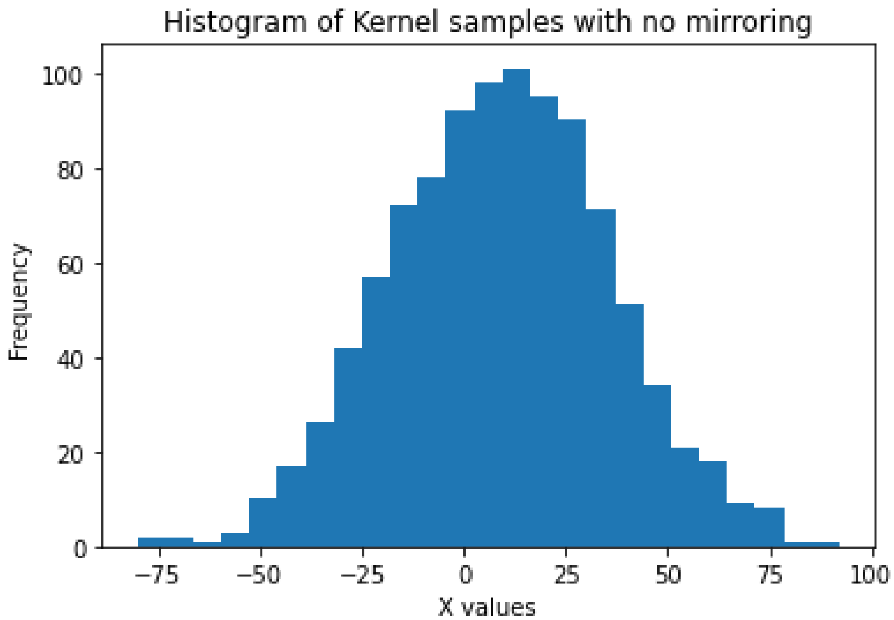
x-value: 8.8125
maximum density value (peak): 0.015

6.3

Using the values found for the Gaussian Kernel used above, we find the following Histogram:



The histogram here seems to differ from the KDE given in part b. The density seems to drop off significantly in the 50's compared to the 20's value range for the histogram in this segment, while in section b they seem to be similar. This could be explained by the mirroring technique we applied generating artificial positive values in this 20's range that overestimates their density. This can be seen if we generate a histogram with no mirroring, seen below:

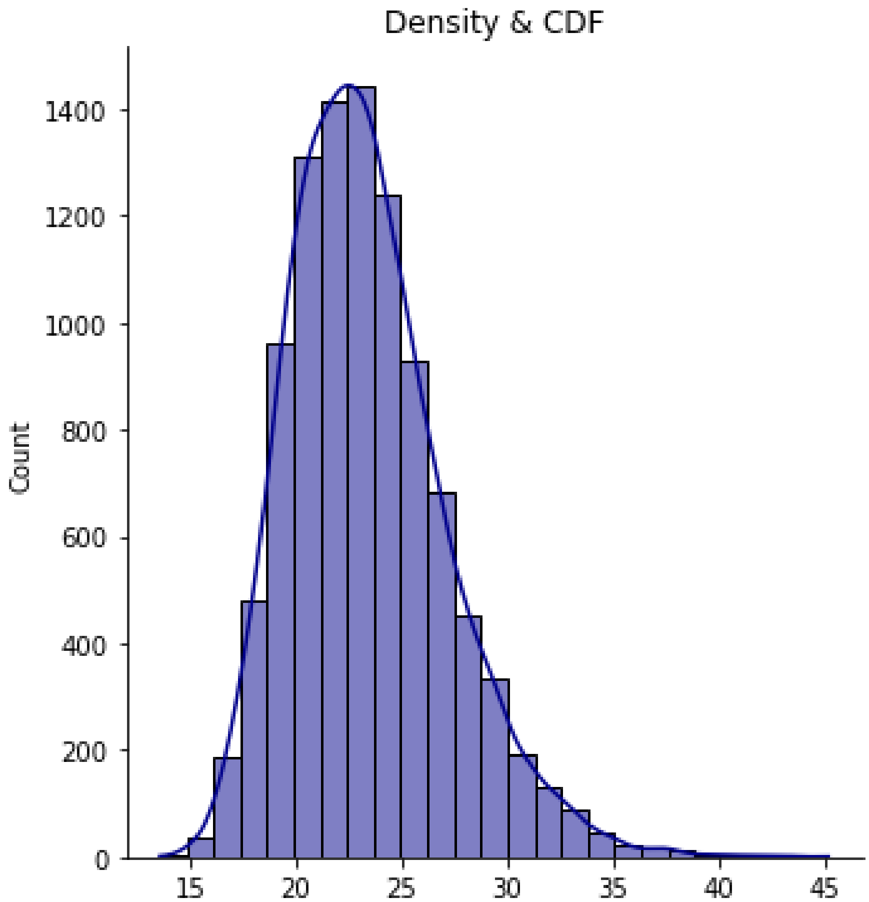


As seen here, there are quite a few negative values, with higher density in the [-20, 0) range, which would explain the discrepancy seen.

7

Please see the attached ipynb for the full solution. The histogram and values for mu and sigma are:

```
mu = 2.6288210511608163
sigma = 2.6288210511608163
```



8

8.1

The general form of the density function for a beta distribution is:

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Given the density function in this case which is:

$$f(x; p) = p(1-x)^{p-1}$$

We can compare the two and see that:

$$p = \beta$$

and

$$x^{\alpha-1} = 1$$
$$\alpha - 1 = 0$$
$$\alpha = 1$$

The mean of a beta distribution is:

$$\mu = \frac{\alpha}{\alpha + \beta}$$

Therefore, in this case, μ is equal to:

$$\frac{1}{1+p}$$

Using the method of moments, we equate this amount to the average of the $\{X_1, X_2, ...X_n\}$ observations and solve for p:

$$\frac{1}{1+p} = \frac{\sum_{i=1}^{n} X_i}{n}$$
$$1 + p = \frac{n}{\sum_{i=1}^{n} X_i}$$
$$p = \frac{n}{\sum_{i=1}^{n} X_i} - 1$$

8.2

The likelihood function is

$$L(p) = \prod_{i=1}^{n} f(X_i; p)$$

The log-likelihood is given by:

$$\begin{aligned} l(p) &= \sum_{i=1}^{n} ln(p(1-x)^{p-1}) \\ &= \sum_{i=1}^{n} (ln(p) + (p-1)ln(1-x_i)) \end{aligned} \tag{7}$$

Deriving the log-likelihood function:

$$\frac{d}{dp} \sum_{i=1}^{n} (ln(p) + (p-1)ln(1-x_i)) = \sum_{i=1}^{n} (\frac{1}{p} + ln(1-X_i))$$

Setting to 0 and solving for p:

$$\sum_{i=1}^{n}(\frac{1}{p} + ln(1 - X_i)) = 0$$

$$\frac{n}{p} = -\sum_{i=1}^{n} ln(1 - X_i)$$

$$\frac{n}{p} = \sum_{i=1}^{n} ln(\frac{1}{1 - X_i})$$

$$p = \frac{n}{\sum_{i=1}^{n} ln(\frac{1}{1-X_i})}$$

Taking the second derivative of the log-likelihood, to verify that the estimator maximizes it:

$$\frac{d}{dp}\sum_{i=1}^{n}(\frac{1}{p} + ln(1 - X_i)) = \sum \frac{-1}{p^2}$$

Since the second derivative is a summation of negative terms, it will always be negative. This implies concavity of the log-likelihood function, indicating that the estimator maximizes the log-likelihood function.

8.3

Please see the attached ipynb for a more detailed approach. In this case, we reject the null hypothesis. The constructed table will also be attached as an Excel spreadsheet.