STAT3035/8035
Tutorial 2

Marco Li

Contact: qingyue.li@anu.edu.au

Outline

Statistics - Estimation

- MOM
- Solve system:

$$E_\theta(X) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\vdots$$
$$E_\theta\left(X^k\right) = \overline{x^k} = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

  where k is number of parameters.

- MOP
- Solve system:

$$x_{p_1} = \hat{x}_{p_1}$$
$$\vdots$$
$$x_{p_k} = \hat{x}_{p_k}$$

  where $\hat{x}_p$ is observed p^{th} percentile of data, for some choice of $p_1, \ldots, p_k$

Statistics - Estimation

- MLE
- Likelihood Function: $L\left(\theta; x_1, \ldots, x_n\right) = \prod_{i=1}^{n} f_X\left(x_i; \theta\right)$
- Log-Likelihood Function: $l(\theta) = \ln\left\{L\left(\theta; x_1, \ldots, x_n\right)\right\}$
- Maximum Likelihood Estimate, $\hat{\theta}_{MLE}$, solves:

$$\frac{\partial l(\theta)}{\partial \theta_i} = 0, \quad 1 \leq i \leq k$$

- Maximum Likelihood Theorem: For large samples,

$$\Pr_\theta \left\{ \frac{\hat{\theta}_{MLE} - \theta}{\sqrt{I^{-1}\left(\hat{\theta}_{MLE}\right)}} \leq t \right\} \approx \Phi(t)$$

  where $\Phi(\cdot)$ is the standard normal CDF and $I(\theta) = -E_\theta\left\{l''(\theta)\right\}$

- So, we can use $\hat{\theta}_{MLE} \pm 1.96\sqrt{I^{-1}\left(\hat{\theta}_{MLE}\right)}$ as an approximate 95% confidence interval for θ

Statistics - Goodness of fit testing

- Pearson Chi-Squared Test
- Idea:
  - Data is n iid observations classified into k categories
  - O_i = number of observations in category i
  - "Theory": $p_i = \Pr(\text{ obs. in cat. } i)$
  - $E_i = np_i$ = expected # of obs. in ith category
  - Measure discrepancy using test statistic:

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

    which has an approximate χ^2 -distribution with a number of degrees of freedom equal to:
    $df = k - 1 - (\#$ parameters estimated in determining $p_i)$
- Implementation:
  - For continuous data, need to "discretise" using bins
  - Choose 5 to 15 bins
  - Could use histogram bins (equal width)

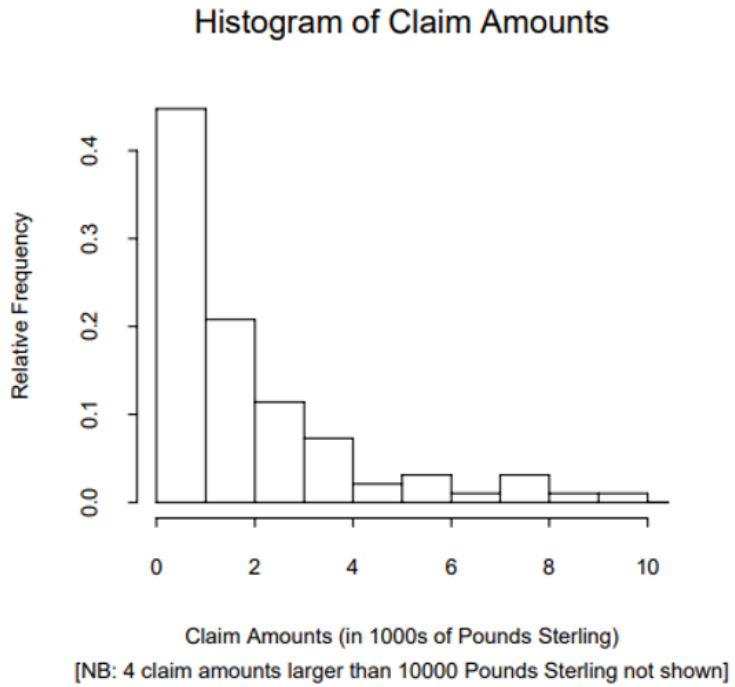
Outline

Question 1

(MOM, MOP, Pearson Chi-square test) This exercise continues the investigation of the example claims dataset from the reading brick. Recall that the observed claims had an average of $\bar{x} = 2989.83$ and sample variance of $s^2 = (6856.11)^2$. In addition, the observed claims had a median of $\hat{x}_{0.5} = 1233.5$, and the upper and lower quantiles of the observed claims are $\hat{x}_{0.75} = 2836.75$ and $\hat{x}_{0.25} = 401$, respectively. We continue our investigation of an exponential model for these data:

Claim Amounts (in £'s)

24	26	73	84	102	115
132	159	207	240	241	254
268	272	282	300	302	329
346	359	367	375	378	384
452	475	495	503	531	543
563	594	609	671	687	691
716	757	821	829	885	893
968	1053	1081	1083	1150	1205
1262	1270	1351	1385	1498	1546
1565	1635	1671	1706	1820	1829
1855	1873	1914	2030	2066	2240
2413	2421	2521	2586	2727	2797
2850	2989	3110	3166	3383	3443
3512	3515	3531	4068	4527	5006
5065	5481	6046	7003	7245	7477
8738	9197	16370	17605	25318	58524



Question 1

(a) We saw that the usual MOM estimate of the exponential mean parameter, θ, in this case was $\hat{\theta}_{MOM} = \bar{x} = 2989.83$. Suppose we chose to modify our MOM procedure so that we matched second moments instead of first moments, what would the new estimate of θ be?
(b) We also saw that the MOP estimate of the exponential mean parameter, θ, in this case was $\hat{\theta}_{MOP} = \hat{x}_{0.5}/\ln(2) = 1779.56$. Suppose we chose to modify our MOP procedure so that we matched upper quantiles instead of medians, what would the new estimate of θ be? What if we chose to match lower quantiles?

Solution 1

Solution 1

(a) The second moment of an exponential random variable with mean parameter θ is $E\left(X^2\right) = 2\theta^2$. Also, the second moment of the data is $n^{-1}\sum_{i=1}^{n} X_i^2 = n^{-1}\left\{(n-1)s^2 + n\bar{x}^2\right\} = 96^{-1}\left\{95(6856.11)^2 + 96(2989.83)^2\right\} = 55455679.38$. Thus, the new estimate is

$$\hat{\theta} = \sqrt{0.5(55455679.38)} = 5265.72.$$

(b) The upper quantile of an exponential distribution with mean parameter θ is the value $x_{0.75}$ which solves the equation:

$$0.75 = \int_0^{x_{0.75}} \theta^{-1} e^{-x/\theta} dx \quad \Longrightarrow \quad x_{0.75} = \theta \ln(4)$$

So, the new estimate based on matching upper quantiles is $\hat{\theta} = \hat{x}_{0.75}/\ln(4)$ =2046.28. Similarly, for the lower quantile we have $x_{0.25} = \theta \ln(4/3)$, so that the new estimate based on matching lower quantiles is $\hat{\theta} = \hat{x}_{0.25}/\ln(4/3) = 1393.90$.

Question 1

(MOM, MOP, Pearson Chi-square test) This exercise continues the investigation of the example claims dataset from the reading brick. Recall that the observed claims had an average of $\bar{x} = 2989.83$ and sample variance of $s^2 = (6856.11)^2$. In addition, the observed claims had a median of $\hat{x}_{0.5} = 1233.5$, and the upper and lower quantiles of the observed claims are $\hat{x}_{0.75} = 2836.75$ and $\hat{x}_{0.25} = 401$, respectively. We continue our investigation of an exponential model for these data:
(c)* For each of the three new estimates you calculated in parts a and b, construct 12 "equal-count" bins and perform Pearson goodness-of-fit tests to decide whether the exponential distribution is an adequate model for the observed counts.
(d)*Repeat part c, this time using 8 "equal-count" bins. Does your opinion regarding the adequacy of the exponential model change? What if you use 6 "equal-count" bins? Discuss.

Solution 1

Solution 1

(c) As in the reading brick, the bin upper endpoints $b_1, \ldots, b_{11}$ can be calculated recursively by letting $b_0 = 0$ and calculating:

$$b_i = -\hat{\theta} \ln \left\{ \exp\left(-\frac{b_{i-1}}{\hat{\theta}} \right) - \frac{8}{96} \right\}, \quad 1 \leq i \leq 11$$

and then recalling that $b_{12} = \infty$. Using this recursion and the estimate from part a, we see that the appropriate bins and their associated observed counts are:

Bin Range	O_i	E_i	Bin Range	O_i	E_i
$0 - 458.18$	25	8	$3649.92 - 4609.97$	2	8
$458.18 - 960.05$	17	8	$4609.97 - 5784.98$	3	8
$960.05 - 1514.85$	11	8	$5784.98 - 7299.84$	3	8
$1514.85 - 2135.07$	12	8	$7299.84 - 9434.90$	3	8
$2135.07 - 2838.20$	7	8	$9434.90 - 13084.82$	0	8
$2838.20 - 3649.92$	9	8	$13084.82+$	4	8

The value of the Pearson test statistic is $X^2 = 73.5$, which has an associated p -value o $\Pr\left(\chi^2_{10} \geq 73.5\right) = 9.3 \times 10^{-12}$. So, we clearly do not think that this is an adequate model for the data.

Solution 1

(c) For the estimate based on the upper quantile matching estimate, we would arrive at bins and a count table of:

Bin Range	O_i	E_i	Bin Range	O_i	E_i
$0 - 178.05$	8	8	$1418.37 - 1791.45$	6	8
$178.05 - 373.08$	13	8	$1791.45 - 2248.07$	8	8
$373.08 - 588.68$	10	8	$2248.07 - 2836.75$	6	8
$588.68 - 829.70$	9	8	$2836.75 - 3666.44$	9	8
$829.70 - 1102.94$	6	8	$3666.44 - 5084.81$	4	8
$1102.94 - 1418.37$	6	8	$13084.82+$	11	8

This table yields a Pearson test statistic of $X^2 = 9$ which has an associated p -value o $\Pr\left(\chi^2_{10} > 9\right) = 0.5321$. Thus, this test seems to suggest that the exponential model under this parameter estimate is adequately modelling the data.

Solution 1

(c) Finally, using the estimate based on matching lower quantiles, we have a count table of:

Bin Range	O_i	E_i	Bin Range	O_i	E_i
$0 - 121.29$	6	8	$966.18 - 1220.32$	6	8
$121.29 - 254.14$	6	8	$1220.32 - 1531.36$	5	8
$254.14 - 401.00$	12	8	$1531.36 - 1932.36$	10	8
$401.00 - 565.18$	7	8	$1932.36 - 2497.53$	5	8
$565.18 - 751.31$	6	8	$2497.53 - 3463.71$	10	8
$751.31 - 966.18$	5	8	$5084.81+$	18	8

This table leads to a Pearson test statistic of $X^2 = 21$ which has an associated p -value of $\Pr\left(\chi^2_{10} > 21\right) = 0.0211$. So, we are back to rejecting the adequacy of the exponential model.

I'll leave part (d) to yourself.

Question 2

(MLE, Pearson Chi-square test) For a certain insurance portfolio, the individual claim amounts have a distribution with density: $f_X(x;\gamma) = \gamma x^{\gamma-1}(1+x)^{-\gamma-1}$. Suppose that we observed the following claim amount data on this portfolio (the claim amounts have been ordered for convenience):

3	3	7	9	12	13	14	16	17	17
18	19	23	26	30	30	32	32	45	57
63	76	77	80	90	109	120	127	149	163
164	174	188	195	203	205	244	257	267	298
308	314	338	393	405	422	534	659	757	1059
1072	1171	1200	1677	2214	2779	3053	15334	17761	50284

Some useful summary statistics are: $\sum_{i=1}^{60} \ln(X_i) = 300.29$ and $\sum_{i=1}^{60} \ln(X_i+1) = 302.03$

(a) Show that the CDF of the claim amount distribution is given by:

$$F_X(x) = \int_0^x f_X(u)du = \left(\frac{x}{1+x}\right)^{\gamma}$$

[HINT: Make a change of integration variable from u to $v = \frac{u}{1+u}$.]

Solution 2

Solution 2

(a) Using the hint, we have:

$$v = \frac{u}{1+u} \quad \Longrightarrow \quad u = \frac{v}{1-v} \quad \Longrightarrow \quad du = \left(\frac{1}{1-v} + \frac{v}{(1-v)^2} \right) dv = \frac{dv}{(1-v)^2}$$

Thus, the CDF is:

$$\begin{aligned}
F_X(x) &= \int_0^x \frac{k u^{k-1}}{(1+u)^{k+1}} du \\
&= \int_0^{x/(1+x)} k \left(\frac{v}{1-v} \right)^{k-1} \left(1 + \frac{v}{1-v} \right)^{-k-1} \frac{dv}{(1-v)^2} \\
&= \int_0^{x/(1+x)} k v^{k-1} (1-v)^{-k+1} (1-v)^{k+1} (1-v)^{-2} dv \\
&= \int_0^{x/(1+x)} k v^{k-1} dv \\
&= v^k \Big|_0^{x/(1+x)} \\
&= \left(\frac{x}{1+x} \right)^k
\end{aligned}$$

Question 2

(b)* Find the method of percentiles estimate of γ using the median.

Question 2

(b)* Find the method of percentiles estimate of γ using the median.
(b) The theoretical median of the distribution of the X_i 's is the value $x_{0.5}$ which solves the equation:

$$\left(\frac{x_{0.5}}{1+x_{0.5}}\right)^k = 0.5 \quad \Longrightarrow \quad x_{0.5} = \frac{\sqrt[4]{0.5}}{1-\sqrt[4]{0.5}} = \frac{1}{\sqrt[4]{2}-1}$$

Furthermore, the median of the observed claim amounts is 163.5. Therefore, the MOP estimate of k is the solution to the equation:

$$\frac{1}{\sqrt[\psi]{2}-1} = 163.5 \quad \Longrightarrow \quad \sqrt[k]{2} = \frac{163.5+1}{163.5} \quad \Longrightarrow \quad \frac{1}{k}\ln 2 = \ln 1.00612$$

which yields $\hat{k}_{MOP} = \frac{\ln 2}{\text{li } 1.00612} = 113.7$.

Question 2

(MLE, Pearson Chi-square test) For a certain insurance portfolio, the individual claim amounts have a distribution with density: $f_X(x;\gamma) = \gamma x^{\gamma-1}(1+x)^{-\gamma-1}$. Suppose that we observed the following claim amount data on this portfolio (the claim amounts have been ordered for convenience):

3	3	7	9	12	13	14	16	17	17
18	19	23	26	30	30	32	32	45	57
63	76	77	80	90	109	120	127	149	163
164	174	188	195	203	205	244	257	267	298
308	314	338	393	405	422	534	659	757	1059
1072	1171	1200	1677	2214	2779	3053	15334	17761	50284

Some useful summary statistics are: $\sum_{i=1}^{60} \ln(X_i) = 300.29$ and $\sum_{i=1}^{60} \ln(X_i + 1) = 302.03$

(c) Find the MLE of λ, as well as an approximate 95% confidence interval for λ.

Solution 2

Solution 2

(c) The log-likelihood for the sample is given by:

$$l(k) = \sum_{i=1}^{n} \ln f_X(x_i) = n \ln k + (k-1) \sum_{i=1}^{n} \ln(X_i) - (k+1) \sum_{i=1}^{n} \ln(1+X_i)$$

Thus, the score equation is:

$$l'(k) = nk^{-1} + \sum_{i=1}^{n} \ln(X_i) - \sum_{i=1}^{n} \ln(1+X_i) = 0$$

which implies that the MLE of k is

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

So, for the given data, the MLE is $\hat{k} = 60(302.03 - 300.29)^{-1} = 34.48$. To find a confidence interval we note that:

$$l''(k) = -nk^{-2} \quad \Longrightarrow \quad I(k) = -E\left\{l''(k)\right\} = nk^{-2}$$

Therefore, the approximate variance of the MLE is given by $\hat{k}^2 n^{-1} = \left(34.48^2\right)/60 = 19.81$. Thus, an approximate 95% confidence interval is $34.48 \pm 1.96\sqrt{19.81} = (25.76, 43.20)$.

Question 2

(d)* Suppose that we choose to test the adequacy of the chosen family of distributions by using a Pearson goodness-of-fit test. Conduct this test employing the MOP estimate of γ from part b and a discretisation based on 5 equal-count bins.

Question 2

(d)* Suppose that we choose to test the adequacy of the chosen family of distributions by using a Pearson goodness-of-fit test. Conduct this test employing the MOP estimate of γ from part b and a discretisation based on 5 equal-count bins.
(d) To create 5 equal-count bins, we need to solve the equation $\Pr\left(b_{i-1} < X_i \leq b_i\right) = \frac{1}{5}$ for b_i. In other words, we have:

$$0.2 = \left(\frac{b_i}{b_i+1}\right)^{\hat{\gamma}} - \left(\frac{b_{i-1}}{b_{i-1}+1}\right)^{\hat{\gamma}} \Longrightarrow \frac{b_i}{b_i+1} = \sqrt[3]{0.2 + \left(\frac{b_{i-1}}{b_{i-1}+1}\right)^{\hat{\gamma}}}$$

$$\Longrightarrow \quad b_i = \frac{\sqrt[4]{0.2 + b_{i-1}^{j}\left(b_{i-1}+1\right)^{-\hat{\gamma}}}}{1 - \sqrt[3]{0.2 + b_{i-1}^{\hat{\gamma}}\left(b_{i-1}+1\right)^{-\hat{\gamma}}}}$$

This yields the 5 bins:

$$(0, 70.15), (70.15, 123.59), (123.59, 222.08), (222.08, 509.04), (509.04, \infty)$$

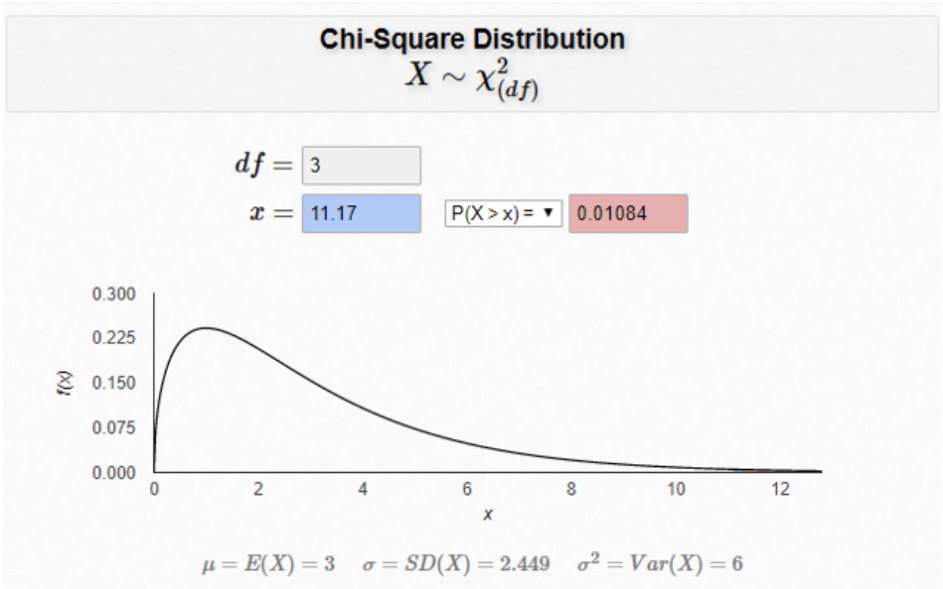
Based on these bins: the E_i 's are all $60/5 = 12$ and the O_i 's are:

$$O_1 = 21, \quad O_2 = 6, \quad O_3 = 9, \quad O_4 = 10, \quad O_5 = 14$$

Solution 2

(d) Therefore, the Pearson test statistic is:

$$\begin{aligned} X^2 &= \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(21-12)^2}{12} + \frac{(6-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(10-12)^2}{12} + \frac{(14-12)^2}{12} \\ &= 11.17 \end{aligned}$$



This test statistic has $5 - 1 - 1 = 3$ degrees of freedom, so an appropriate comparison value would be $\chi^2_{(3)}(0.95) = 7.81$ in which case we would reject the null hypothesis at significance level $\alpha = 0.05$ that the chosen family is an adequate model for the observations. However, $\chi^2_{(3)}(0.99) = 11.34$, and we would now not reject the null hypothesis at a significance of $\alpha = 0.01$.

Question 3*

The data below is a sample of 100 individual claim amounts made on an insurance portfolio:

2.25	2.99	3.15	3.17	3.30	3.35	3.38	3.39	3.40	3.51
3.55	3.55	3.77	3.85	3.98	3.99	4.08	4.11	4.19	4.27
4.41	4.54	4.55	4.90	4.95	5.21	5.24	5.49	5.73	5.78
5.85	5.88	5.95	5.95	5.95	5.97	6.29	6.33	6.40	6.57
6.60	6.63	6.69	6.72	6.88	6.95	6.99	7.10	7.14	7.36
7.45	7.74	7.84	8.05	8.42	8.49	8.56	8.66	8.74	8.76
8.80	8.93	9.10	9.29	9.29	9.29	9.39	9.40	9.42	9.56
9.63	9.64	9.77	9.90	9.92	9.99	10.18	10.48	11.12	11.29
11.34	11.34	11.40	11.63	11.64	11.65	12.36	12.62	13.11	14.05
14.23	14.50	14.93	15.06	15.70	16.02	16.35	17.03	23.84	26.33

Simple calculations show that summary statistics on the sample are:

$$\sum_{i=1}^{100} \ln X_i = 198.6413, \quad \sum_{i=1}^{100} \{\ln X_i\}^2 = 419.4468$$

and the sample quantiles are $\hat{x}_{0.25} = 5.014666$ and $\hat{x}_{0.75} = 9.972587$.

(a)* Suppose we choose to model this claims data using the lognormal family. Find the method of percentile (MOP) estimates of μ and σ^2 based on quantiles.

Solution 3

Solution 3

(a) We first need to find the quartiles of the lognormal distribution, $x_{0.25}$ and $x_{0.75}$

$$\begin{aligned} 0.25 &= \Pr\left(X \leq x_{0.25}\right) = \Pr\left(\tfrac{\ln X-\mu}{\sigma} \leq \tfrac{\ln x_{0.25}-\mu}{\sigma}\right) = \Phi\left(\tfrac{\ln x_{0.25}-\mu}{\sigma}\right) \\ 0.75 &= \Pr\left(X \leq x_{0.75}\right) = \Pr\left(\tfrac{\ln X-\mu}{\sigma} \leq \tfrac{\ln x_{0.75}-\mu}{\sigma}\right) = \Phi\left(\tfrac{\ln x_{0.75}-\mu}{\sigma}\right) \end{aligned}$$

which implies that $x_{0.25} = \exp\left\{\mu + \sigma\Phi^{-1}(0.25)\right\} = \exp\{\mu - 0.6745\sigma\}$ and $x_{0.75} = \exp\{\mu + \sigma\Phi^{-1}(0.75)\} = \exp\{\mu + 0.6745\sigma\}$. Therefore, the MOP equations are:

$$\begin{aligned} 5.0147 &= \exp\{\mu - 0.6745\sigma\} \\ 9.9726 &= \exp\{\mu + 0.6745\sigma\} \end{aligned}$$

These equations have solutions: $\hat{\mu} = 1.9561$ and $\hat{\sigma}^2 = 0.2597$.

Question 3

(b)* Compare the MOP estimates to the MLEs. Does the lognormal model seem plausible?

Lecture slides Section 2.4.2, slide 3: Maximum Likelihood Estimate (MLE) of Log-Likelihood Function: Can use "equivariance"

μ, σ^2 are mean, variance for normal data $y_i = \ln(x_i)$

$$\hat{\mu}_{MLE} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} \ln(x_i) = \overline{\ln(x)}$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} \{y_i - \bar{y}\}^2 = \frac{1}{n}\sum_{i=1}^{n} \left\{\ln(x_i) - \overline{\ln(x)}\right\}^2$$

$$= \overline{\ln(x)^2} - \overline{\ln(x)}^2$$

Question 3

(b)* Compare the MOP estimates to the MLEs. Does the lognormal model seem plausible?

Lecture slides Section 2.4.2, slide 3: Maximum Likelihood Estimate (MLE) of Log-Likelihood Function: Can use "equivariance"

μ, σ^2 are mean, variance for normal data $y_i = \ln(x_i)$

$$\hat{\mu}_{MLE} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} \ln(x_i) = \overline{\ln(x)}$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} \{y_i - \bar{y}\}^2 = \frac{1}{n}\sum_{i=1}^{n} \left\{\ln(x_i) - \overline{\ln(x)}\right\}^2$$

$$= \overline{\ln(x)^2} - \overline{\ln(x)}^2$$

(b) The MLEs are $\hat{\mu} = 1.9864$ and $\hat{\sigma}^2 = 4.1945 - \left(1.9864^2\right) = 0.2487$. The similarity between the MOP estimates and the MLEs suggests that the lognormal model may be reasonable.

Question 3

(c)* Construct 10 “equal count” bins and test of the adequacy of the lognormal model.

Question 3

(c)* Construct 10 "equal count" bins and test of the adequacy of the lognormal model.

(c) Let the bin endpoints be denoted $b_0, b_1, \ldots, b_{10}$. Clearly, $b_0 = 0$ and $b_{10} = \infty$, and the other endpoints are calculated based on the equation:

$$\frac{10}{100} = \Pr\left(b_{i-1} < X \leq b_i\right) = \Pr\left(\frac{\ln b_{i-1} - \mu}{\sigma} < \frac{\ln X - \mu}{\sigma} \leq \frac{\ln b_i - \mu}{\sigma}\right)$$
$$= \Phi\left(\frac{\ln b_i - \mu}{\sigma}\right) - \Phi\left(\frac{\ln b_{i-1} - \mu}{\sigma}\right)$$

which implies the recursion relationship:

$$b_i = \exp\left[\mu + \sigma\Phi^{-1}\left\{0.1 + \Phi\left(\frac{\ln b_{i-1} - \mu}{\sigma}\right)\right\}\right]$$

Using this relationship, and employing the MOP estimates to replace μ and σ, we have: $b_1 = 3.68, b_2 = 4.60, b_3 = 5.41, b_4 = 6.21, b_5 = 7.07, b_6 = 8.05, b_7 = 9.24, b_8 = 10.86, b_9 = 13.59$. Therefore, the observed counts in each of the ten bins are: $O_1 = 12, O_2 = 11, O_3 = 4, O_4 = 9, O_5 = 11, O_6 = 6, O_7 = 10, O_8 = 15, O_9 = 11, O_{10} = 11$. Since all the E_i' s are clearly 10, this leads to a Pearson statistic of $X^2 = 8.6$. Since this test statistic has $10 - 1 - 2 = 7$ degrees of freedom, the p -value is $\Pr\left(\chi^2_{(T)} > 8.6\right) = 0.2827$ and we see that the lognormal model is indeed reasonable for these data.