ACSC/STAT 4703, Actuarial Models II
Fall 2016

Toby Kenney

Homework Sheet 5
Model Solutions

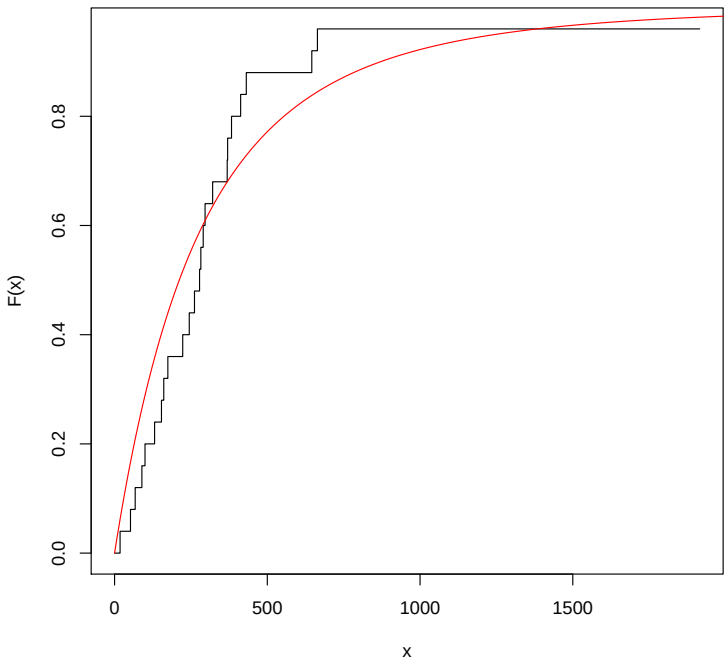
Basic Questions

1. *An insurance company is modelling claim data as following a Pareto distribution with $\alpha = 4$. It collects the following sample of claims:*

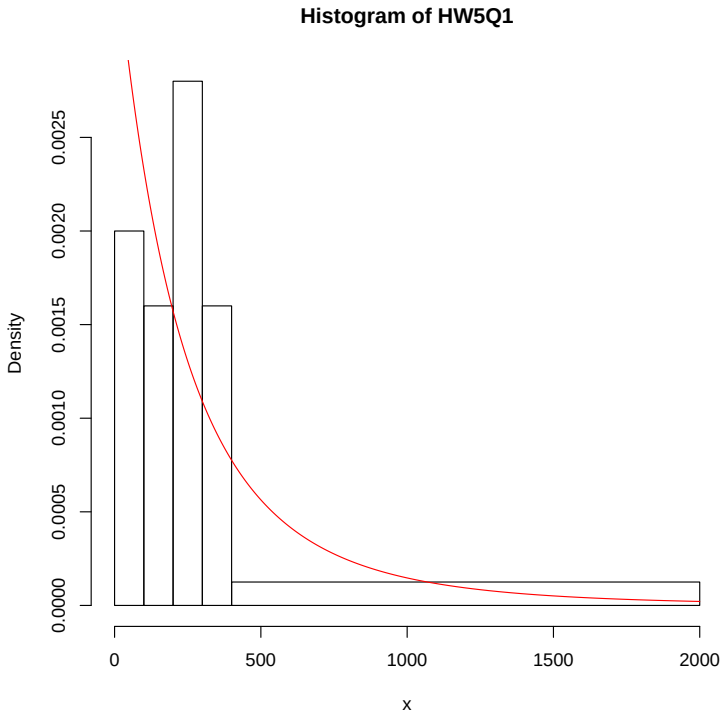
```
18.0  52.1  67.5  89.4  99.6  131.0  153.5  161.0  174.4  223.1
244.5  261.6  278.2  282.4  290.1  296.2  321.0  368.7  370.1
382.8  412.7  431.2  645.6  664.0  1915.5
```

The MLE for θ is 1119.3399. Graphically compare this empirical distribution with the best fitting Pareto distribution with $\alpha = 4$. Include the following plots:

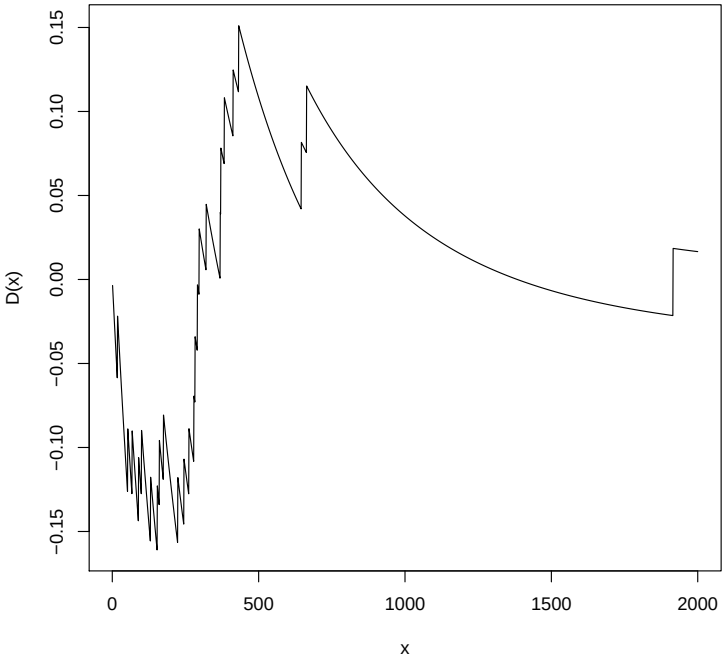
(a) Comparisons of $F(x)$ and $F^*(x)$



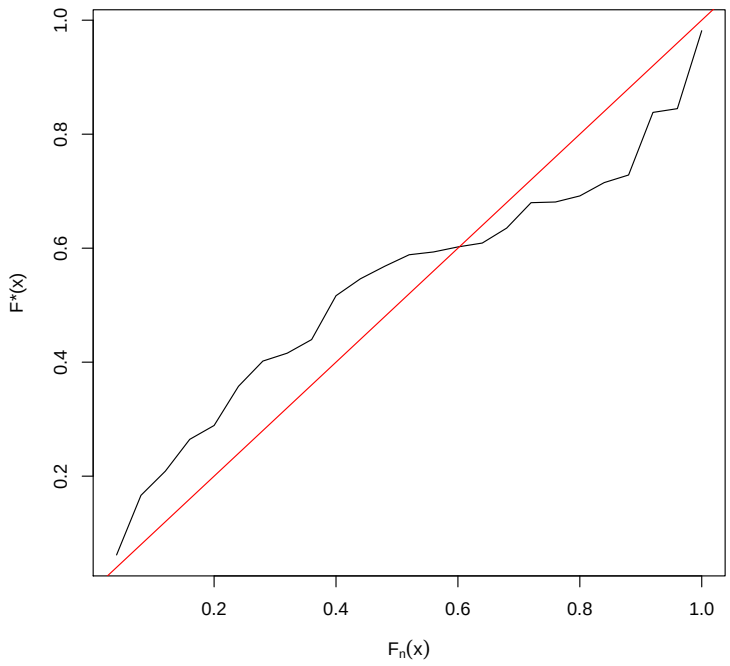
(b) Comparisons of $f(x)$ and $f^*(x)$



(c) A plot of $D(x)$ against x.



(d) A p-p plot of $F(x)$ against $F^*(x)$.



2. *For the data in Question 1, calculate the following test statistics for the goodness of fit of the Pareto distribution with* $\alpha = 4$ *and* $\theta = 1119.3399$ *using:*

   (a) The Kolmogorov-Smirnov test.

   The Kolmogorov-Smirnov test statistic is the maximum value of $|D(x)|$. We look at the following table:

x	$F^*(x)$	$F_n(x)$	$D(x^-)$	$D(x^+)$
18.0	0.0618	0.04	−0.0618	−0.0218
52.1	0.1664	0.08	−0.1264	−0.0864
67.5	0.2088	0.12	−0.1288	−0.0888
89.4	0.2646	0.16	−0.1446	−0.1046
99.6	0.2889	0.20	−0.1289	−0.0889
131.0	0.3577	0.24	−0.1577	−0.1177
153.5	0.4019	0.28	−0.1619	−0.1219
161.0	0.4158	0.32	−0.1358	−0.0958
174.4	0.4396	0.36	−0.1196	−0.0796
223.1	0.5166	0.40	−0.1566	−0.1166
244.5	0.5463	0.44	−0.1463	−0.1063
261.6	0.5683	0.48	−0.1283	−0.0883
278.2	0.5885	0.52	−0.1085	−0.0685
282.4	0.5934	0.56	−0.0734	−0.0334
290.1	0.6022	0.60	−0.0422	−0.0022
296.2	0.6090	0.64	−0.0090	0.0310
321.0	0.6353	0.68	0.0047	0.0447
368.7	0.6798	0.72	0.0002	0.0402
370.1	0.6810	0.76	0.0390	0.0790
382.8	0.6917	0.80	0.0683	0.1083
412.7	0.7151	0.84	0.0849	0.1249
431.2	0.7284	0.88	0.1116	0.1516
645.6	0.8382	0.92	0.0418	0.0818
664.0	0.8448	0.96	0.0752	0.1152
1915.5	0.9815	1.00	−0.0215	0.0185

From this we see that the Kolmogorov-Smirnov statistic is 0.1619.

(b) The Anderson-Darling test.

The Anderson-Darling test statistic can be computed as

$$A^2 = -n + n\sum_{j=0}^{k}(1 - F_n(y_j))^2 \left(\log(1 - F^*(y_j)) - \log(1 - F^*(y_{j+1}))\right) + n\sum_{j=0}^{k}(F_n(y_j))^2 \left(\log(F^*(y_j)) - \log\right.$$

y_j	$F_n(y_j)$	$F^*(y_j)$	$(1-F(y_j))^2$ $(\log(1-F^*(y_j)) - \log(1-F^*(y_{j+1})))$	$F(y_j)^2$ $(\log(F^*(y_{j+1})) - \log(F^*(y_j)))$
0.0	0	0	0.6919	—
18.0	0.0618	0.04	0.1089	0.0016
52.1	0.1664	0.08	0.0442	0.0015
67.5	0.2088	0.12	0.0566	0.0034
89.4	0.2646	0.16	0.0237	0.0022
99.6	0.2889	0.20	0.0651	0.0085
131.0	0.3577	0.24	0.0412	0.0067
153.5	0.4019	0.28	0.0122	0.0027
161.0	0.4158	0.32	0.0193	0.0057
174.4	0.4396	0.36	0.0605	0.0209
223.1	0.5166	0.40	0.0228	0.0089
244.5	0.5463	0.44	0.0156	0.0077
261.6	0.5683	0.48	0.0129	0.0080
278.2	0.5885	0.52	0.0028	0.0022
282.4	0.5934	0.56	0.0042	0.0046
290.1	0.6022	0.60	0.0028	0.0040
296.2	0.6090	0.64	0.0090	0.0173
321.0	0.6353	0.68	0.0133	0.0314
368.7	0.6798	0.72	0.0003	0.0009
370.1	0.6810	0.76	0.0020	0.0090
382.8	0.6917	0.80	0.0032	0.0213
412.7	0.7151	0.84	0.0012	0.0131
431.2	0.7284	0.88	0.0075	0.1087
645.6	0.8382	0.92	0.0003	0.0066
664.0	0.8448	0.96	0.0034	0.1382
1915.5	0.9815	1.00	0.0000	0.0187
			0.5738048	0.4538695

The Anderson-Darling statistic is then $25(0.5738048 + 0.4538695 - 1) = 0.6918572$. The critical value at the 95% confidence level is 2.492, so the statistic is not significant.

(c) The chi-square test, dividing into the intervals 0–200, 200–400, and more than 400.

We obtain the following table:

Interval	Expected	Observed	$\frac{(E-O)^2}{E}$
$[0, 200]$	12.04727	9	0.770785
$[200, 400]$	5.587803	11	5.24211
$[400, \infty)$	7.364922	5	0.7593911

So the chi-square statistic is $0.770785 + 5.24211 + 0.7593911 = 6.772286$.

3. *For the data in Question 1, perform a likelihood ratio test to determine whether a Pareto distribution with fixed $\alpha = 4$, or a Pareto distribution with α freely estimated is a better fit for the data. [The MLE for the general Pareto distribution is $\alpha = 22.49267$ and $\theta = 7159.3127$.]*

The log-likelihood of the data point x_i under the Pareto distribution is $\log(\alpha) + \alpha \log(\theta) - (\alpha + 1) \log(\theta + x_i)$. The log-likelihood of the data is therefore

$$25 \log(\alpha) + 25\alpha \log(\theta) - (\alpha + 1) \sum \log(\theta + x_i)$$

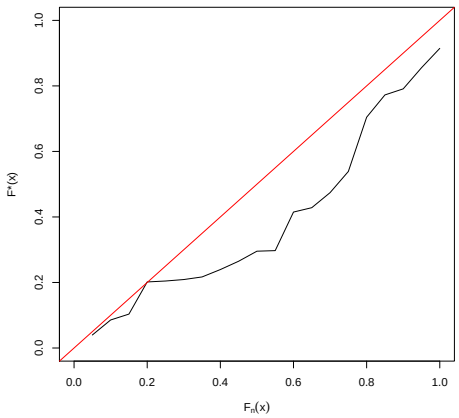
For $\alpha = 4, \theta = 1119.3399$, this log-likelihood is -170.7396, while for $\alpha = 22.49267$ and $\theta = 7159.3127$, the log-likelihood is -170.186. The log-likelihood ratio statistic is twice the difference between these, or

$$2(-170.186 - (-170.7396)) = 1.1072$$

This is smaller than the critical value for a chi-square distribution with one degree of freedom. Therefore, a Pareto distribution with freely chosen α does not fit the data significantly better than a Pareto distribution with $\alpha = 4$.

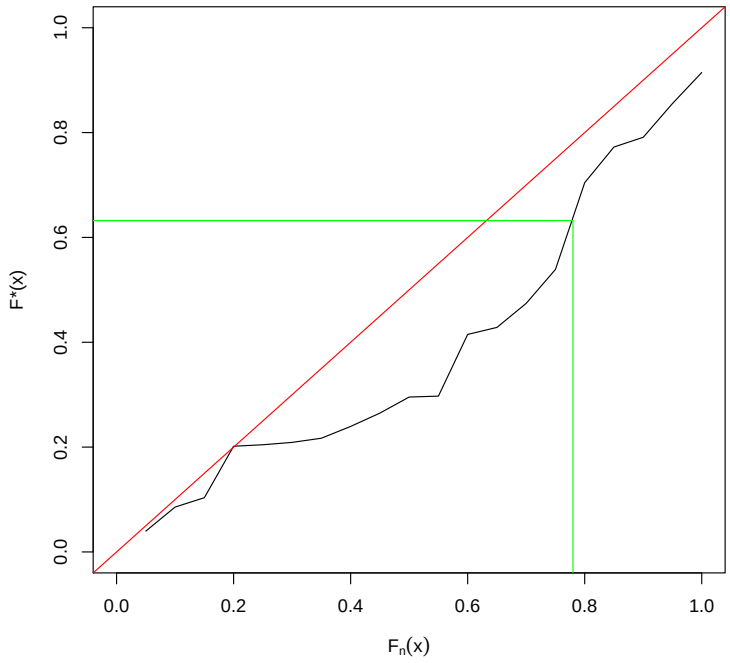
Standard Questions

4. *An insurance company collects a sample of 20 past claims, and attempts to fit a distribution to the claims. Based on experience with other claims, the company believes that a Weibull distribution with $\tau = 2$ and $\theta = 2,400$ may be appropriate to model these claims. It constructs the following p-p plot to compare the sample to this distribution:*



(a) How many of the points in their sample were less than 2,400?

Under the Weibull distribution they are fitting, the distribution function has $F(2400) = 1 - e^{-\left(\frac{1}{1}\right)^2} = 1 - e^{-1} = 0.6321206$. Looking at the graph



we see that the corresponding $F_n(x)$ is about 0.78, which means that there are $20 \times 0.78 = 15$ samples less than 2400.

(b) Which of the following statements best describes the fit of the Weibull distribution to the data:

(i) The Weibull distribution assigns too much probability to high values and too little probability to low values.

(ii) The Weibull distribution assigns too much probability to low values and too little probability to high values.

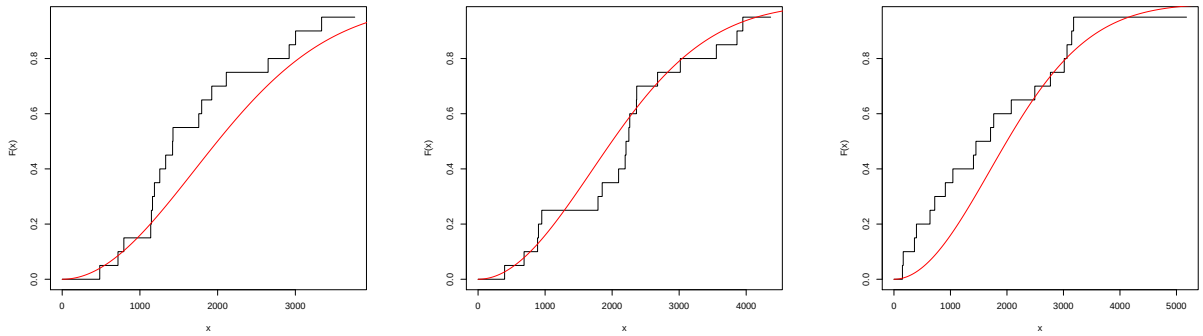
(iii) The Weibull distribution assigns too much probability to tail values and too little probability to central values.

(iv) The Weibull distribution assigns too much probability to central values and too little probability to tail values.

We see from the graph that $F^*(x) < F_n(x)$ for nearly all values of x. This means that the Weibull distribution assigns too much probability to high values, and too little probability to low values.

(c) Which of the following plots shows the empirical distribution function? Justify your answer.

(i) (ii) (iii)



As we noted in part (b), we have $F_n(x) > F^*(x)$ for almost all x, and particularly for larger values of x. This is true for graph (i), but for graph (ii) there are a number of values of x with $F_n(x) < F^*(x)$. For graph (iii) we mostly have $F_n(x) > F^*(x)$, but there are a number of values around $F_n(x) = 0.7$ where they are close, whereas in the p-p plot, we see that when $F_n(x) = 0.7$, we have $F^*(x) \approx (0.5)$, so they are not close. Therefore, (i) must be the corresponding plot.