ACSC/STAT 4703, Actuarial Models II
Fall 2015

Toby Kenney

Homework Sheet 5
Model Solutions

Basic Questions

1. *An insurance company is modelling claim data as following a Weibull distribution with $\tau = 3$. It collects the following sample of claims:*

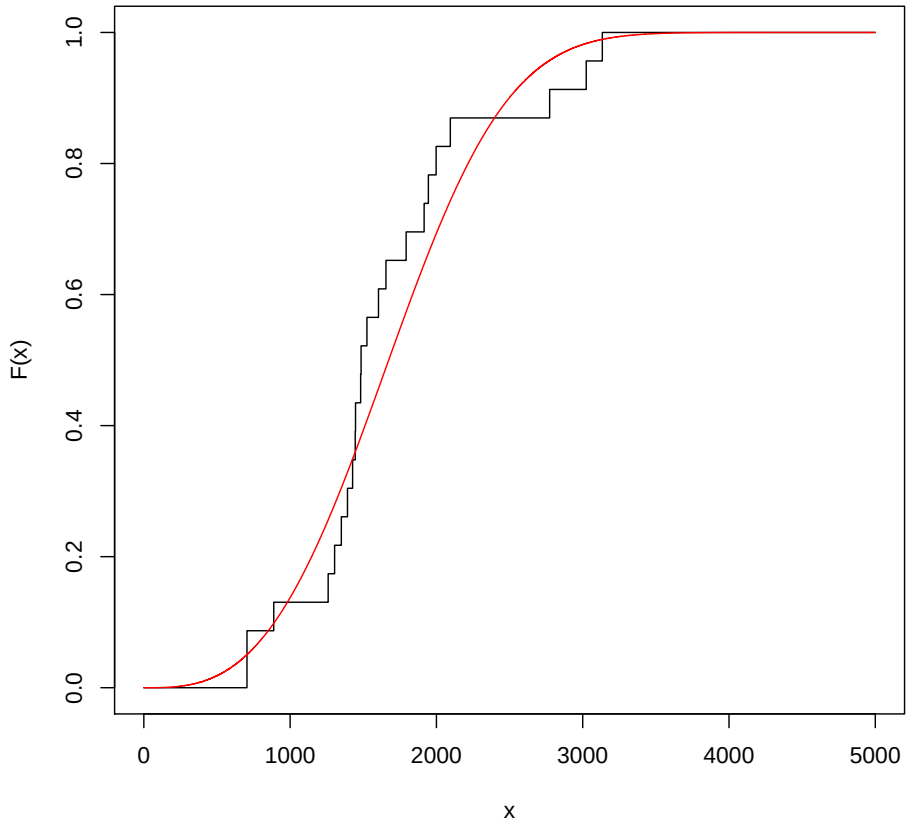
   1,445 1,392 1,655 1,260 1,525 1,604 3,134 2,095 1,447
   1,304 1,350 1,793 1,945 888 1,485 1,998 1,916 2,774
   1,482 705 1,427 705 3,024

   The MLE for θ is

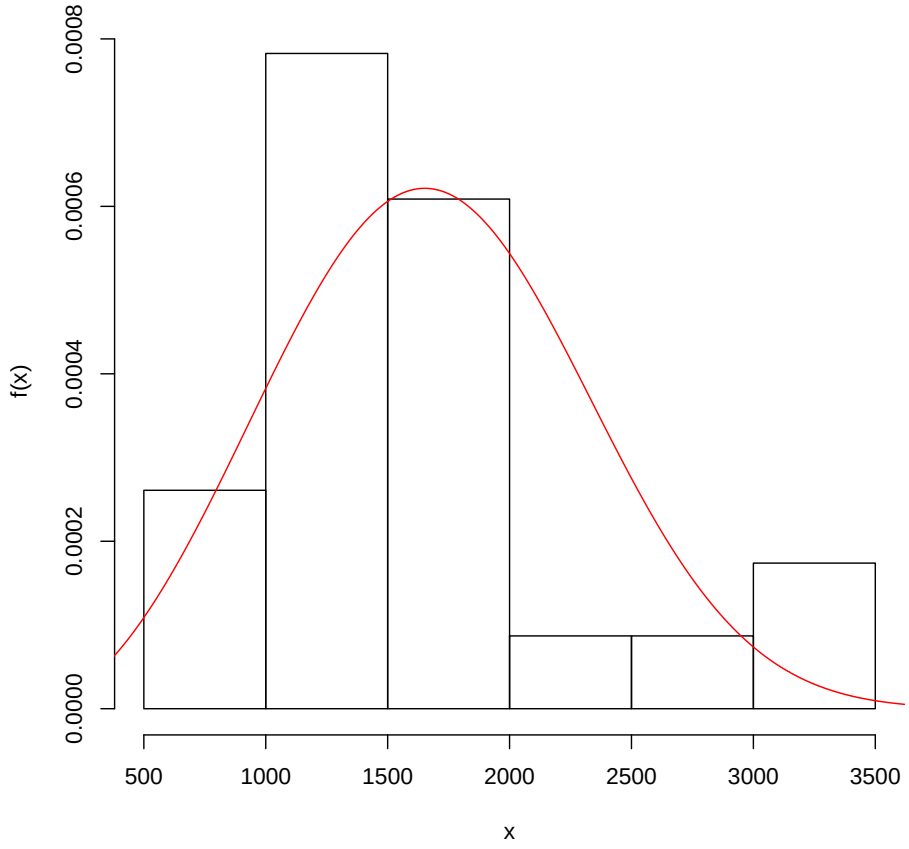
   $$\left(\frac{\sum_{i=1}^{23} {X_i}^3}{23}\right)^{\frac{1}{3}} = 1891.194$$

   Graphically compare this empirical distribution with the best fitting Weibull distribution with $\tau = 3$. Include the following plots:

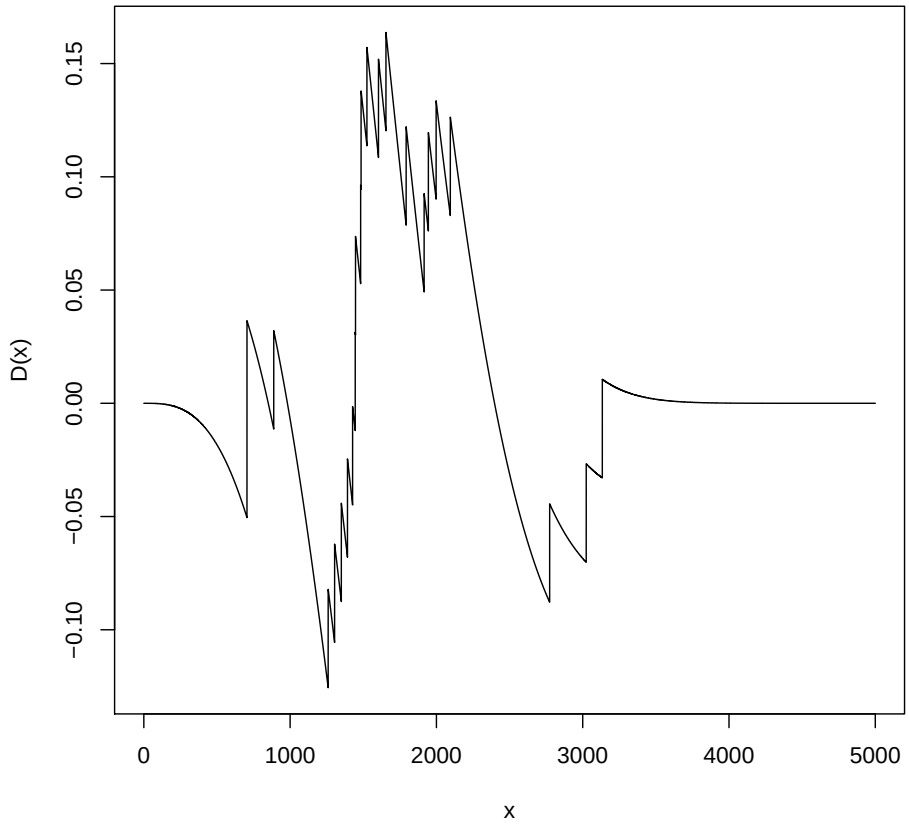
   (a) Comparisons of $F(x)$ and $F^*(x)$



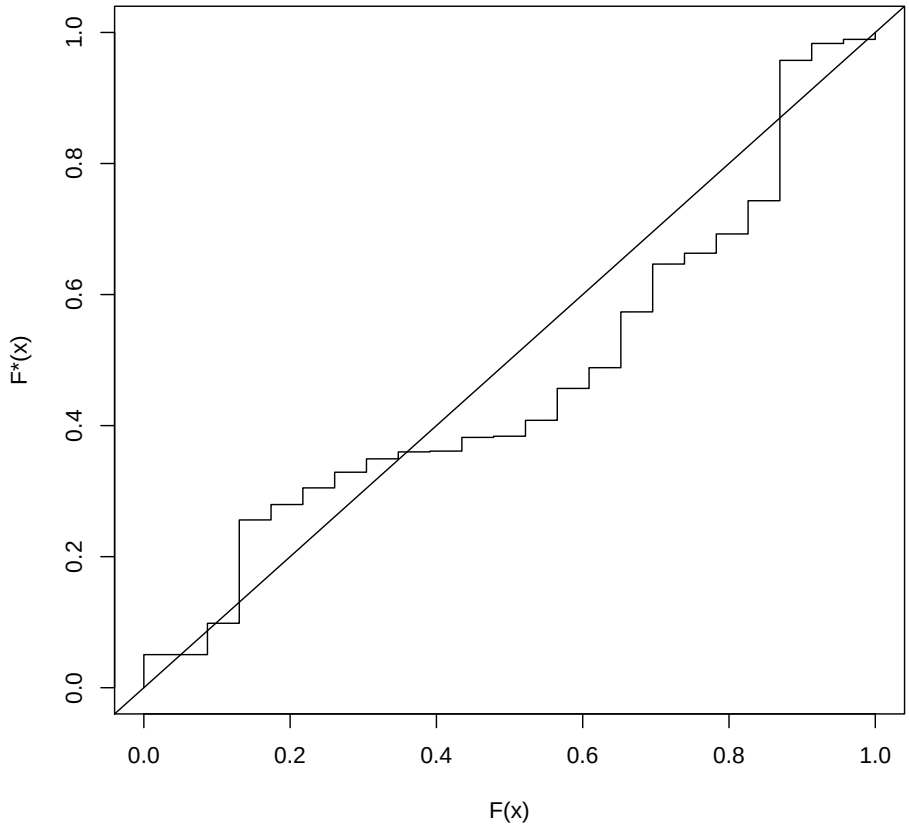
(b) Comparisons of $f(x)$ and $f^*(x)$



(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 Weibull distribution with* $\tau = 3$ *and* $\theta = 1891.194$ *using:*

   (a) The Kolmogorov-Smirnov test.

   The Kolmogorov-Smirnov test statistic is the maximum absolute value of $D(x)$. We compute the following table

x	$F(x^-)$	$F(x)$	$F^*(x)$	$D(x^-)$	$D(x)$
705	0.00000000	0.08695652	0.05048457	−0.05048457	0.036471950
888	0.08695652	0.13043478	0.09834338	−0.01138686	0.032091399
1260	0.13043478	0.17391304	0.25601578	−0.12558100	−0.082102736
1304	0.17391304	0.21739130	0.27950137	−0.10558832	−0.062110062
1350	0.21739130	0.26086957	0.30492927	−0.08753797	−0.044059709
1392	0.26086957	0.30434783	0.32884693	−0.06797737	−0.024499106
1427	0.30434783	0.34782609	0.34923027	−0.04488244	−0.001404178
1445	0.34782609	0.39130435	0.35985596	−0.01202988	0.031448384
1447	0.39130435	0.43478261	0.36104216	0.03026219	0.073740453
1482	0.43478261	0.47826087	0.38196606	0.05281655	0.096294806
1485	0.47826087	0.52173913	0.38377318	0.09448769	0.137965951
1525	0.52173913	0.56521739	0.40804572	0.11369341	0.157171675
1604	0.56521739	0.60869565	0.45670638	0.10851101	0.151989271
1655	0.60869565	0.65217391	0.48837901	0.12031664	0.163794902
1793	0.65217391	0.69565217	0.57351689	0.07865702	0.122135279
1916	0.69565217	0.73913043	0.64649871	0.04915347	0.092631728
1945	0.73913043	0.78260870	0.66304430	0.07608614	0.119564398
1998	0.78260870	0.82608696	0.69246761	0.09014109	0.133619351
2095	0.82608696	0.86956522	0.74318242	0.08290453	0.126382792
2774	0.86956522	0.91304348	0.95739621	−0.08783099	−0.044352730
3024	0.91304348	0.95652174	0.98323140	−0.07018792	−0.026709656
3134	0.95652174	1.00000000	0.98944135	−0.03291962	0.010558645

We see that the maximum value of $|D(x)|$ is 0.163794902, when $x = 1,655$.

The critical value at the 95% level is $\frac{1.358}{\sqrt{23}} = 0.2831626$, so this test statistic is not significant.

(b) The Anderson-Darling test.

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

j	y_j	$F_n(y_j)$	$F^*(y_j)$	$(1-F(y_j))^2$ $(\log(1-F^*(y_j)) - \log(1-F^*(y_{j+1})))$	$F(y_j)^2$ $(\log(F^*(y_{j+1})) - \log(F$
0	0	0	0	0.0518035	—
1	705	0.08695652	0.05048457	0.0431146465	0.0050419470
2	888	0.13043478	0.09834338	0.1453413459	0.0162778153
3	1260	0.17391304	0.25601578	0.0218895028	0.0026546118
4	1304	0.21739130	0.27950137	0.0220062232	0.0041149659
5	1350	0.26086957	0.30492927	0.0191299284	0.0051388488
6	1392	0.30434783	0.32884693	0.0149251197	0.0055705405
7	1427	0.34782609	0.34923027	0.0070020777	0.0036261459
8	1445	0.39130435	0.35985596	0.0006871951	0.0005038955
9	1447	0.43478261	0.36104216	0.0106368127	0.0106497240
10	1482	0.47826087	0.38196606	0.0007971065	0.0010796071
11	1485	0.52173913	0.38377318	0.0091918148	0.0166940693
12	1525	0.56521739	0.40804572	0.0162154063	0.0359920397
13	1604	0.60869565	0.45670638	0.0091972229	0.0248431203
14	1655	0.65217391	0.48837901	0.0220202757	0.0683488024
15	1793	0.69565217	0.57351689	0.0173848674	0.0579672086
16	1916	0.73913043	0.64649871	0.0032621627	0.0138056796
17	1945	0.78260870	0.66304430	0.0043181024	0.0265935137
18	1998	0.82608696	0.69246761	0.0054507197	0.0482334771
19	2095	0.86956522	0.74318242	0.0305629579	0.1915128749
20	2774	0.91304348	0.95739621	0.0070505473	0.0221976999
21	3024	0.95652174	0.98323140	0.0008744110	0.0057604292
22	3134	1.00000000	0.98944135	0.0000000000	0.0106147833
				0.4628619	0.5772218

The Anderson-Darling statistic is then $23(0.4628619 + 0.5772218 - 1) = 0.9219251$. 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 0–1500, 1500–2000, and more than 2000.

We have the following:

	O	E	$\frac{(O-E)^2}{E}$
0—1500	12	9.035276	0.972808
1500—2000	9	6.916507	0.6276207
2000—∞	4	7.048218	1.318295
total			2.918724

For a chi-squared distribution with 1 degree of freedom, the critical value at the 95% significance level is 3.841459, so this test is not significant.

3. *For the data in Question 1, perform a likelihood ratio test to determine whether a Weibull distribution with fixed $\tau = 3$, or a Weibull distribution with τ freely estimated is a better fit for the data. [The MLE for the general Weibull distribution is $\tau = 2.3831$ and $\theta = 1785.085$.]*

The log-likelihood is

$\sum_i \left(\log(\tau) - \tau\log(\theta) + (\tau-1)\log(x_i) - \left(\frac{x}{\theta}\right)^\tau\right)$.

For $\tau = 3$ and $\theta = 1891.194$, this is -180.1976.

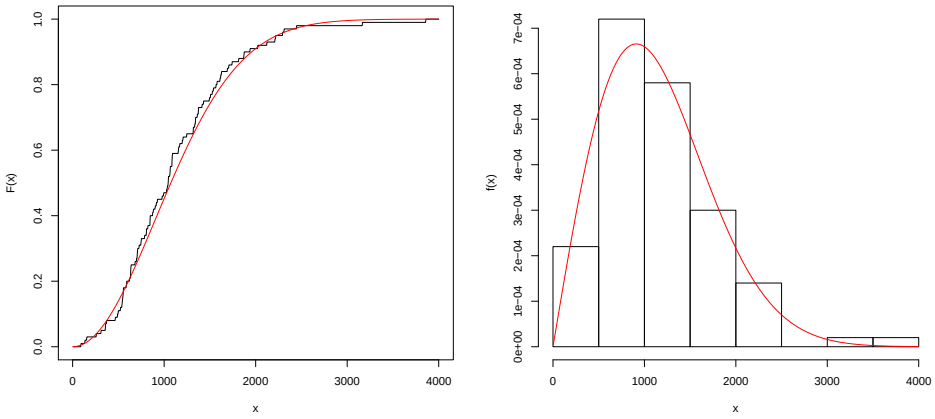
For $\tau = 2.831$ and $\theta = 1872.335$, this is -180.1239.

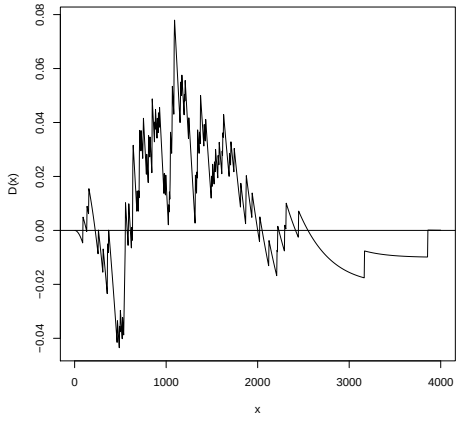
The log-likelihood ratio is therefore $2(-180.1239-(-180.1976)) = 0.1474$. This is not significant.

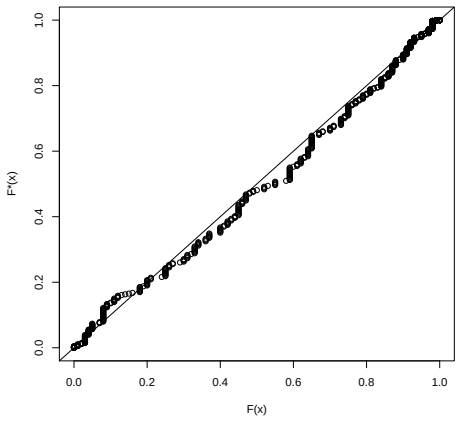
Standard Questions

4. *An insurance company is modelling a data set. It is considering 3 models, each with 1 parameter to be estimated. Below are various diagnostic plots of the fit of each model.*

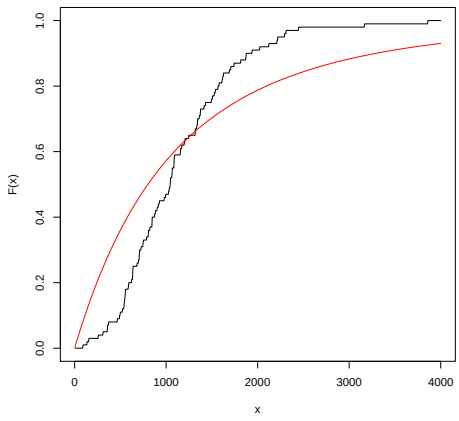
Model I

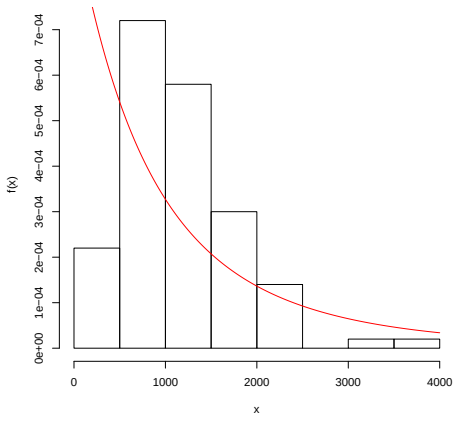


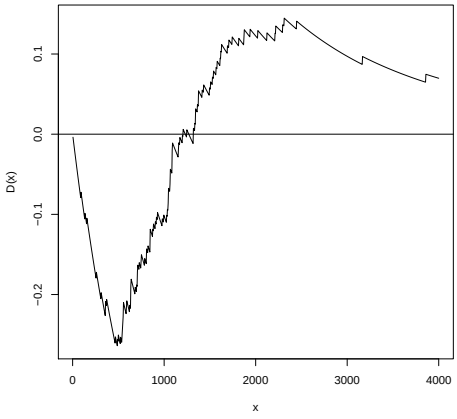

D(x)
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
0
1000
2000
3000
4000
x

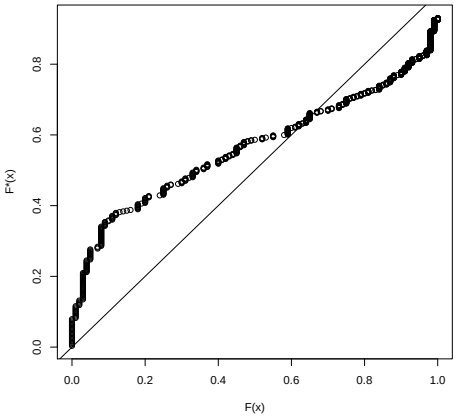

F*(x)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
F(x)

Model II

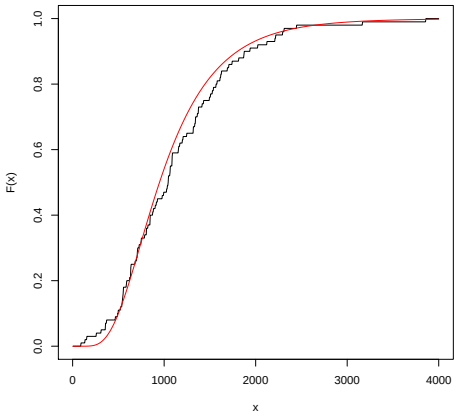

F(x)
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
x

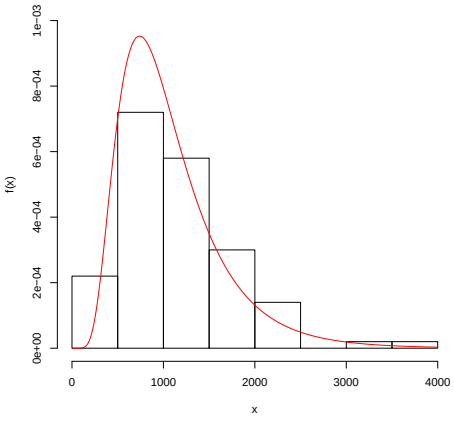

f(x)
7e−04
6e−04
5e−04
4e−04
3e−04
2e−04
1e−04
0e+00
0
1000
2000
3000
4000
x

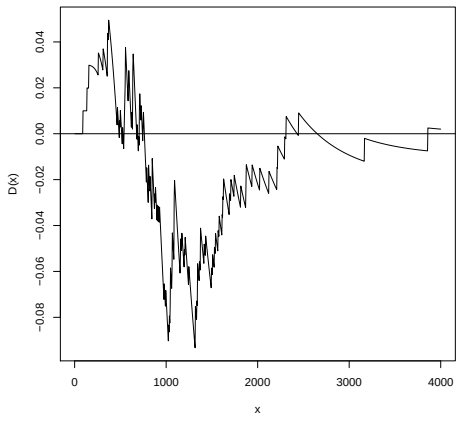

D(x)
0.1
0.0
−0.1
−0.2
0
1000
2000
3000
4000
x

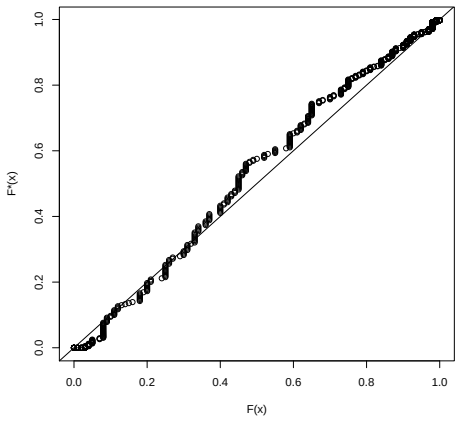

F*(x)
0.0
0.2
0.4
0.6
0.8
0.0
0.2
0.4
0.6
0.8
1.0
F(x)

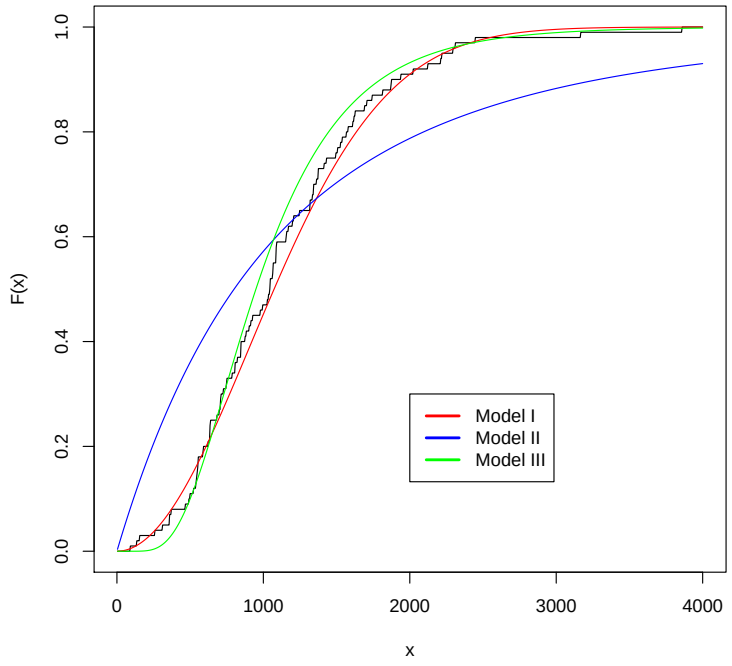
Model III

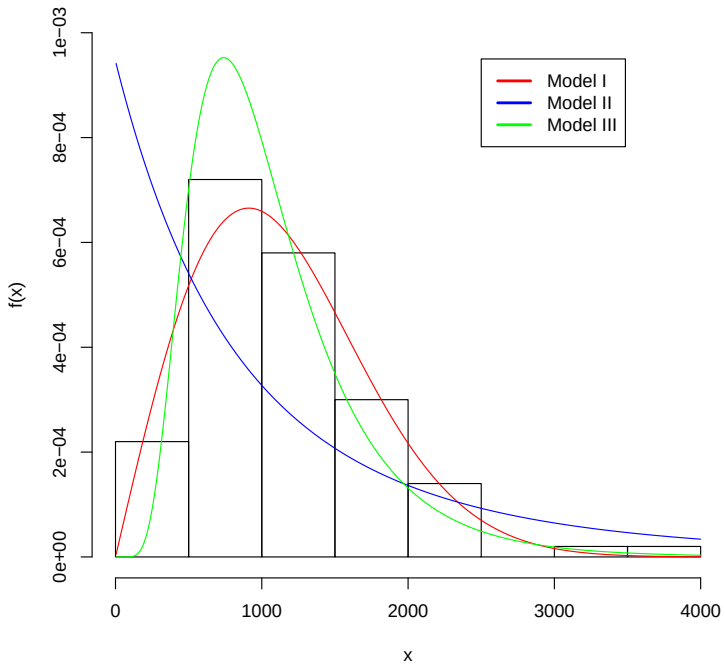

F(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
x


f(x)
0e+00
2e−04
4e−04
6e−04
8e−04
1e−03
0
1000
2000
3000
4000
x


D(x)
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
0
1000
2000
3000
4000
x


F*(x)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
F(x)


F(x)
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
x
Model I
Model II
Model III



Determine which model they should use for the data in the following situations. Justify your answers.

(a) Which model should they choose if accurately estimating (right-hand) tail probabilities is most important, and it is particularly important not to underestimate tail probabilities?

Model I underestimates the tail probabilities, and model II is not a very good fit, so they should choose model III.

(b) The company is considering imposing a deductible, and therefore wants to model the distribution very accurately on small values of x.

We see that model I is very close to the empirical distribution near zero. Therefore, they should choose model I in this situation.

(c) The company uses the Kolmogorov-Smirnov statistic to decide the best model.

The easiest way to find the Kolmogorov-Smirnov statistic is by looking at the plot of $D(x)$. For model I, we see that the largest absolute value of $D(x)$ is slightly less than 0.08; For model II, it is about 0.25, and for model III, it is about 0.09, so model I would be preferred by this measure.