ACSC/STAT 4703, Actuarial Models II
Fall 2017
Toby Kenney
Homework Sheet 2
Model Solutions

Basic Questions

1. *An insurance company has the following portfolio of workers compensation insurance policies:*

Type of worker	*Number*	*Probability of claim*	*mean claim*	*standard deviation*
Manual labourer	*1700*	*0.01*	*\$54,000*	*\$129,000*
Technician	*800*	*0.002*	*\$20,000*	*\$39,000*
Manager	*200*	*0.001*	*\$25,000*	*\$20,000*

Calculate the cost of reinsuring losses above \$10,000,000, if the loading on the reinsurance premium is one standard deviation above the expected claim payment on the reinsurance policy using a gamma approximation for the aggregate losses on this portfolio.

The aggregate losses have mean $1700\times 0.01\times 54000+800\times 0.002\times 20000+200\times 0.001\times 25000=\$955,000$ and variance $1700\times 0.01\times 129000^2+800\times 0.002\times 39000^2+200\times 0.001\times 20000^2+1700\times 0.01\times 0.99\times 54000^2+800\times 0.002\times 0.998\times 20000^2+200\times 0.001\times 0.999\times 25000^2=335,250,475,000$.

Matching moments for the gamma distribution gives

$$\begin{aligned}\alpha\theta &= 955000\\ \alpha\theta^2 &= 335250475000\\ \alpha &= \frac{955000^2}{335250475000} = 2.7204286586\\ \theta &= \frac{335250475000}{955000} = 351047.617801\end{aligned}$$

This means that $\frac{10000000}{\theta}=28.4861639644$. If we denote this by a, then expected claim on the reinsurance is

$$\begin{aligned}\theta\int_a^\infty \frac{(x-a)x^{\alpha-1}e^{-x}}{\Gamma(\alpha)}\,dx &= \theta\left(\int_a^\infty \frac{\alpha x^\alpha e^{-x}}{\Gamma(\alpha+1)}\,dx - a\int_a^\infty \frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)}\right)\\ &= 0.00003403306\end{aligned}$$

The expected square of the reinsurance payment is:

$$\theta^2 \int_a^\infty \frac{(x-a)^2 x^{\alpha-1} e^{-x}}{\Gamma(\alpha)}\, dx = \theta^2 \left(\int_a^\infty \frac{\alpha(\alpha+1) x^{\alpha+1} e^{-x}}{\Gamma(\alpha+2)}\, dx - 2a\alpha \int_a^\infty \frac{x^\alpha e^{-x}}{\Gamma(\alpha+1)}\, dx + a^2 \int_a^\infty \frac{x^{\alpha-1} e^{-x}}{\Gamma(\alpha)}\, d \right.$$
$$= 25.27237$$

The variance of the reinsurance payment is therefore $25.27237 - 0.00003403306^2 = 25.27237$, so the premium for the reinsurance is $0.00003403306 + \sqrt{25.27237} = 5.027197$.

2. *An insurance company is modelling claim data as following a Pareto distribution with $\alpha = 5$. It collects the following sample of claims:*

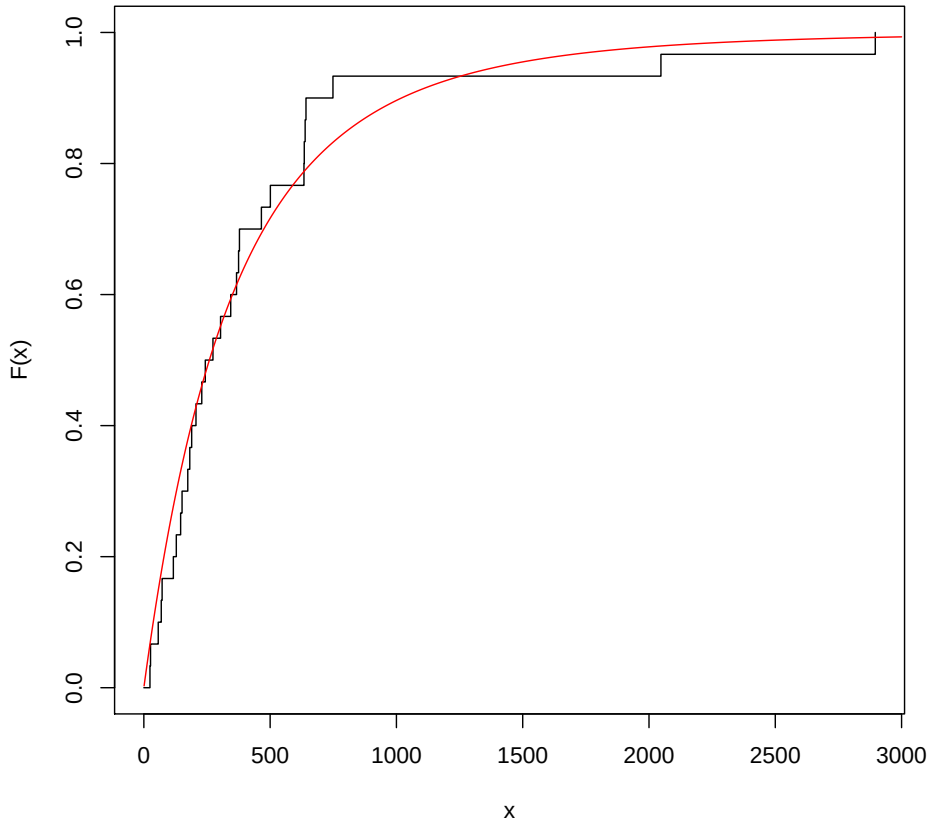
   24.2 26.5 56.9 68.9 72.3 116.8 128.5 145.6 151.3 173.9
   181.8 189.4 206.4 229.3 243.3 273.6 303.7 344.0 367.0
   375.0 378.5 465.4 500.9 633.9 635.1 638.6 641.9 748.7
   2047.2 2895.9

   The MLE for θ is 1744.23679. Graphically compare this empirical distribution with the best fitting Pareto distribution with $\alpha = 5$. Include the following plots:

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

   R code:

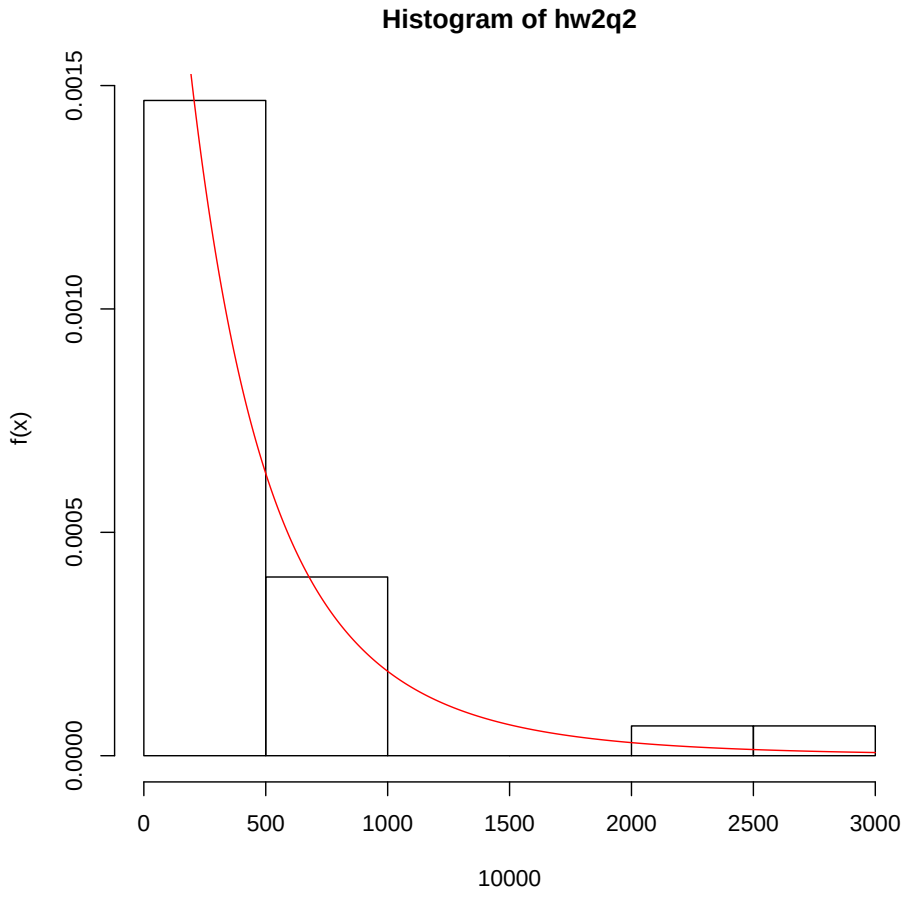
```
plot(c(0,as.vector(rbind(hw2q2,hw2q2))),
    (as.vector(rbind(0:30,0:30))[1:61])/30,type='l',
     xlab="x",ylab="F(x)")
points(1:3000,1-(1744.23679/(1744.23679+1:3000))^5,type='l',col="red")
```



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

R code:

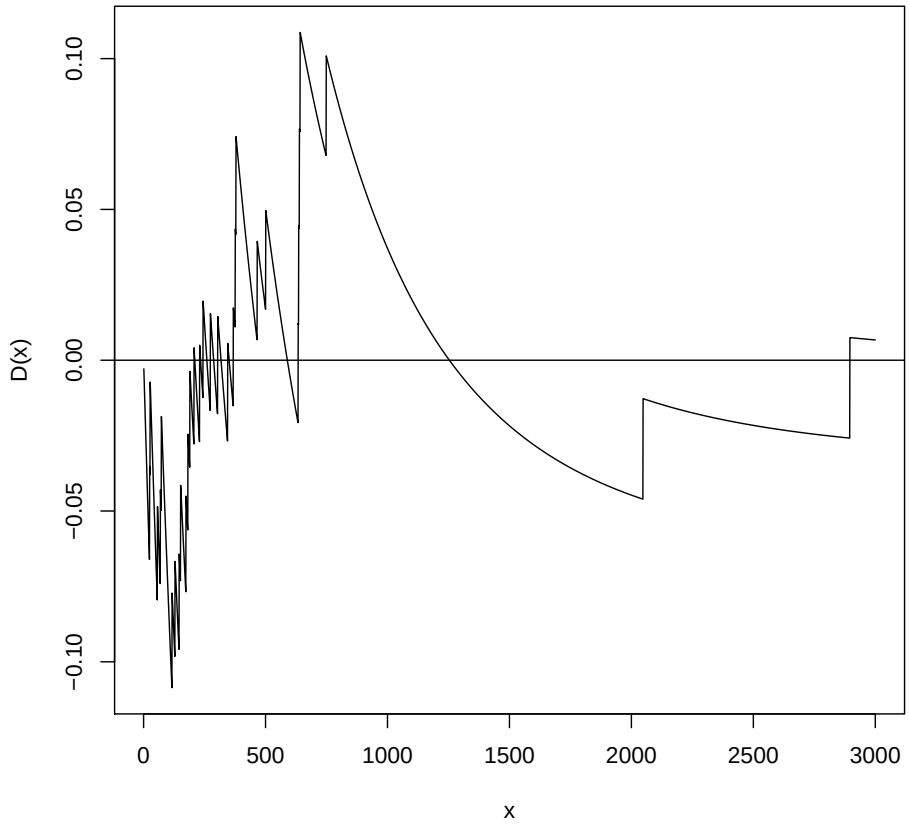
```
hist(hw2q2,probability=T,xlab=x,ylab="f(x)")
points(1:3000,5*1744.23679^5/(1744.23679+1:3000)^6,type='l',col="red")
```



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

R code:

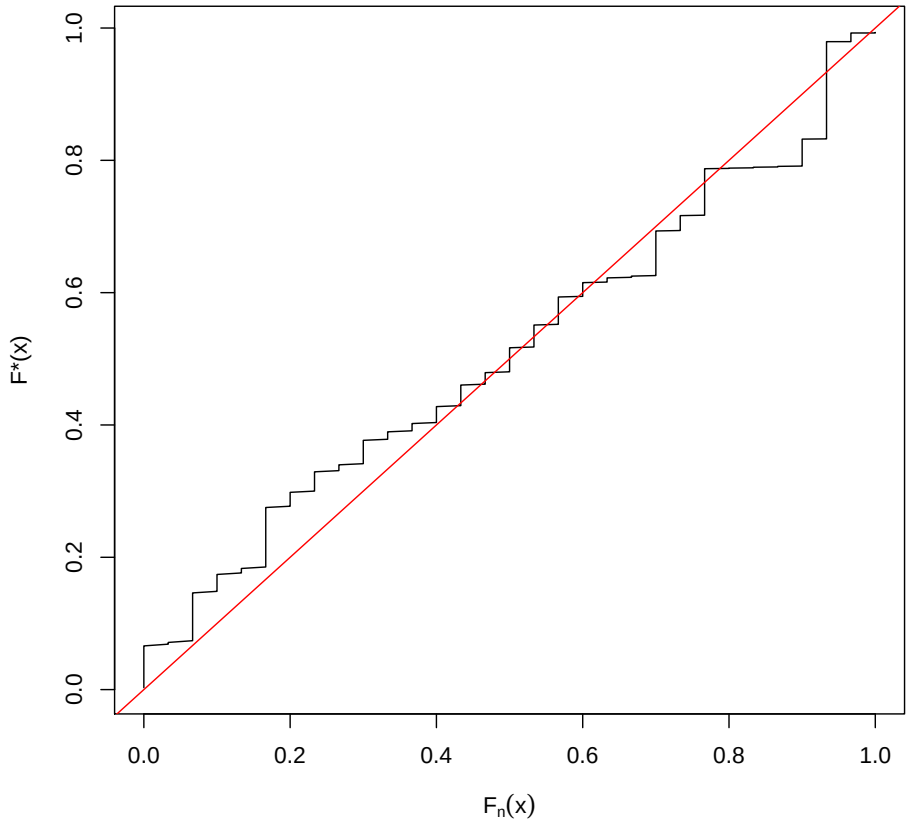
```
hw2q2Fx<-rowSums((rep(1,3000)%*%t(hw2q2))<((1:3000)%*%t(rep(1,30))))/30
hw2q2Fstarx<-1-1744.23679^5/(1744.23679+1:3000)^5
plot(1:3000,hw2q2Fx-hw2q2Fstarx,type='l',xlab="x",ylab="D(x)")
abline(h=0)
```



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

R code:

```
plot(hw2q2Fx,hw2q2Fstarx,type='l',xlab=expression(F[n](x)),ylab="F*(x)")
abline(0,1,col="red")
```



3. *For the data in Question 2, calculate the following test statistics for the goodness of fit of the Pareto distribution with* $\alpha = 5$ *and* $\theta = 1744.23679$*:*

   (a) The Kolmogorov-Smirnov test.

   From the plot of $D(x)$, it is easy to see that the largest absolute values of $D(x)$ occur just before 116.8 and just after 641.9. The corresponding values are:

x	$F_n(x)$	$F^*(x)$	$\lvert D(x)\rvert$
116.8^-	$\frac{5}{30}$	$1 - \left(\frac{1744.23679}{1744.23679+116.8}\right)^5 = 0.276809987586$	0.110143320919
641.9^+	$\frac{27}{30}$	$1 - \left(\frac{1744.23679}{1744.23679+641.9}\right)^5 = 0.791286506023$	0.108713493977

   So the Kolmogorov-Smirnov statistic is 0.108713493977. For a distribution with no estimated parameters, the critical value would be $\frac{1.358}{\sqrt{30}} = 0.247935744364$ so the statistic is not significant.

(b) The Anderson-Darling test.

The Anderson-Darling test statistic for a finite sample is given by

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

For our dataset, we calculate this in the following table:

y_j	$F_n(y_j)$	$F^*(y_j)$	$(1-F_n(y_j))^2$ $(\log(1-F^*(y_j))-\log(1-F^*(y_{j+1})))$	$F_n($ $(\log(F^*(y_{j+1}))-\log(F^*(y$
0.0	0	0	0.0688944825891	
24.2	$\frac{1}{30}$	0.066574832511	0.00607266774222	0.000096589319
26.5	$\frac{2}{30}$	0.072621208235	0.0741415299978	0.003172829
56.9	$\frac{3}{30}$	0.148286116488	0.0268934697426	0.00171909
68.9	$\frac{4}{30}$	0.176100267654	0.00703583523003	0.000759043
72.3	$\frac{5}{30}$	0.183781897108	0.0840344468694	0.01137726
116.8	$\frac{6}{30}$	0.276809987586	0.0200548435718	0.00310057
128.5	$\frac{7}{30}$	0.299120268244	0.0267132827787	0.005392010
145.6	$\frac{8}{30}$	0.330260804256	0.00809784182622	0.002123182
151.3	$\frac{9}{30}$	0.340270167722	0.0290379588141	0.009518776
173.9	$\frac{10}{30}$	0.378230611629	0.00913360435199	0.003654585
181.8	$\frac{11}{30}$	0.390877951332	0.00789820258351	0.004024248
189.4	$\frac{12}{30}$	0.402754742157	0.015755942394	0.00985070
206.4	$\frac{13}{30}$	0.428330321439	0.0187390485379	0.013694910
229.3	$\frac{14}{30}$	0.460736434614	0.0100534330487	0.008676540
243.3	$\frac{15}{30}$	0.479463328255	0.0189124543608	0.019031997
273.6	$\frac{16}{30}$	0.517389204223	0.0161229609665	0.018330015
303.7	$\frac{17}{30}$	0.551828254907	0.0182963402675	0.023342585
344.0	$\frac{18}{30}$	0.59343639487	0.00876309072064	0.012910547
367.0	$\frac{19}{30}$	0.61510480751	0.00254240291849	0.004674408
375.0	$\frac{20}{30}$	0.62231495463	0.000916764260444	0.002210884
378.5	$\frac{21}{30}$	0.625418367445	0.0180548814547	0.050635761
465.4	$\frac{22}{30}$	0.693505453403	0.00566694993422	0.017903929
500.9	$\frac{23}{30}$	0.716982571864	0.0156666255714	0.055328773
633.9	$\frac{24}{30}$	0.787752269611	0.0001008938888	0.00043425
635.1	$\frac{25}{30}$	0.788286957466	0.000204155178611	0.001364400
638.6	$\frac{26}{30}$	0.789837256762	0.000123017406222	0.001376928
641.9	$\frac{27}{30}$	0.791286506023	0.0021892898466	0.040954969
748.7	$\frac{28}{30}$	0.832324010273	0.00931741329541	0.1417398
2047.2	$\frac{29}{30}$	0.979393315918	0.00112221554938	0.012417147
2895.9	$\frac{30}{30}$	0.992494609477	0	0.007533697
			0.530556045699	0.48735047

The Anderson-Darling statistic is therefore $30(0.530556045699+0.487350471547-1) = 0.5371955175$. For a fully specified distribution, the critical value is 2.492: for the Pareto distribution with one parameter estimated, the critical value is even higher, so we cannot reject the Pareto distribution with $\alpha = 5$.

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

We have the following:

interval	Frequency (O)	Expected frequency (E)	$\frac{(O-E)^2}{E}$
0–200	12	12.56577	0.02547362
200–400	9	6.748975	0.7507975
> 400	9	10.68525	0.2657933
			1.042064

The chi-square statistic is 1.042. This is compared to a chi-squared distribution with $3 - 1 - 1 = 1$ degrees of freedom. The critical value at the 95% significance level is 3.841459, so it is not significant.

4. *For the data in Question 2, perform a likelihood ratio test to determine whether a Pareto distribution with fixed $\alpha = 5$, or a Pareto distribution with α freely estimated is a better fit for the data. [The MLE for the general Pareto distribution is $\alpha = 4.641528$ and $\theta = 1599.8973$.]*

   The log-likelihood of a Pareto distribution is

$$l(x) = \sum \log(\alpha) + \alpha \log(\theta) - (\alpha + 1) \log(\theta + x_i)$$

   For $\alpha = 5$, $\theta = 1744.23679$, this is -211.74632212. For the Pareto distribution with $\alpha = 4.641528$ and $\theta = 1599.8973$, the log-likelihood is -211.74291824. The log likelihood statistic is $2(-211.74291824-(-211.74632212)) = 0.00680776$. We compare this to a chi-squared distribution with 1 degree of freedom and conclude that a Pareto distribution with $\alpha = 5$ is the better fit.

5. *For the data in Question 2, use AIC and BIC to choose between a Pareto distribution with $\alpha = 4$ and a gamma distribution for the data. [The MLE for the gamma distribution is $\alpha = 1.021439$ and $\theta = 432.8697$.]*

   In Question 4, we calculated the log-likelihood for the Pareto distribution was -211.74291824. The log-likelihood for a gamma distribution is $\sum(\alpha - 1) \log(x_i) - \frac{x_i}{\theta} - \alpha \log(\theta) - log(\Gamma(\alpha)) = -212.7452$. The gamma distribution has two parameters, whereas the Pareto distribution with α fixed has one parameter. The AIC for the Pareto distribution is therefore $-211.74291824 - 2 \times 1 = -213.74291824$, whereas the AIC for the gamma distribution is $-212.7452 - 2 \times 2 = -216.7452$. There are 30 data points, so the BIC for the Pareto distribution is $-211.74291824 - \frac{1 \times \log(30)}{2} = -213.4435$, while the BIC for the gamma distribution $-212.7452 - \frac{2 \times \log(30)}{2} = -216.1464$. Both AIC and BIC prefer the Pareto distribution.

Standard Questions

6. *An insurance company insures drivers in three provinces and has the following estimates:*

Province	*Probability of claim*	*mean claim*	*standard deviation*
Nova Scotia	*0.08*	*$3,200*	*$5,900*
New Brunswick	*0.03*	*$2,100*	*$3,400*
PEI	*0.02*	*$2,300*	*$4,600*

The insurance company estimates the mean μ and standard deviation σ for the aggregate loss distribution, and buys stop-loss reinsurance for aggregate losses more than $3\mu - \frac{\mu^2}{\sigma}$. The reinsurer models aggregate losses as following a Pareto distribution and sets its premium as 125% of the expected claims on the stop-loss policy. The insurer already insures 2,200 drivers in Nova Scotia, and 980 drivers in New Brunswick. How many drivers should it insure in PEI in order to minimise the reinsurance cost as a proportion of expected claims on the policy?

Let the number of drivers from the three provinces be a, b and c. The expected aggregate losses are then $\mu = 256a + 63b + 46c$. The variance of aggregate losses is $\sigma^2 = 3538464a + 475131b + 526884c$. For a Pareto distribution with parameters α and θ, attachment point of the reinsurance is $a = \frac{\theta}{\alpha-1} + \sqrt{\frac{\theta^2}{(\alpha-1)^2(\alpha-2)}}$, and the expected payment on the reinsurance is

$$\begin{aligned}
\int_a^\infty (x-a)\frac{\alpha\theta^\alpha}{(\theta+x)^{\alpha+1}}\,dx &= \int_a^\infty (\theta + x - (\theta+a))\frac{\alpha\theta^\alpha}{(\theta+x)^{\alpha+1}}\,dx \\
&= \alpha\theta^\alpha\left(\int_a^\infty \frac{1}{(\theta+x)^\alpha}\,dx - (\theta+a)\int_a^\infty \frac{1}{(\theta+x)^{\alpha+1}}\,dx\right) \\
&= \alpha\theta^\alpha\left[-\frac{1}{(\alpha-1)(\theta+x)^{\alpha-1}}\,dx + (\theta+a)\frac{1}{\alpha(\theta+x)^\alpha}\,dx\right]_a^\infty \\
&= \alpha\theta^\alpha\left(\frac{1}{(\alpha-1)(\theta+a)^{\alpha-1}} - (\theta+a)\frac{1}{\alpha(\theta+a)^\alpha}\right) \\
&= \alpha\frac{\theta^\alpha}{(\theta+a)^{\alpha-1}}\left(\frac{1}{(\alpha-1)} - \frac{1}{\alpha}\right) \\
&= \frac{\theta^\alpha}{(\alpha-1)(\theta+a)^{\alpha-1}}
\end{aligned}$$

Since we have

$$\begin{aligned}\frac{\theta}{\alpha-1} &= \mu \\ \frac{\theta^2}{(\alpha-1)^2(\alpha-2)} &= \sigma^2 \\ \alpha &= \frac{\mu^2}{\sigma^2}+2 \\ \theta &= \mu\left(\frac{\mu^2}{\sigma^2}+1\right) \\ a &= 3\mu-\frac{\mu^2}{\sigma}\end{aligned}$$

We get

$$\frac{\theta}{\theta+a} = \frac{\mu^2+\sigma^2}{\mu^2-\mu\sigma+4\sigma^2}$$

The expected payment on the stop-loss insurance is

$$\mu\left(\frac{\mu^2+\sigma^2}{\mu^2-\mu\sigma+4\sigma^2}\right)^{\frac{\mu^2+\sigma^2}{\sigma^2}}$$

We are therefore aiming to minimise

$$\left(\frac{\mu^2+\sigma^2}{\mu^2-\mu\sigma+4\sigma^2}\right)^{\frac{\mu^2+\sigma^2}{\sigma^2}}$$

If we let $r=\frac{\mu}{\sigma}$ then this expression becomes

$$\left(\frac{r^2+1}{r^2-r+4}\right)^{r^2+1}$$

Differentiating the logarithm of this with respect to r, we see that the minimum occurs at a solution to

$$\begin{aligned}2r\log\left(\frac{r^2+1}{r^2-r+4}\right)+(r^2+1)\left(\frac{-r^2+6r+1}{(r^2-r+4)(r^2+1)}\right) &= 0 \\ 2r\log\left(\frac{r^2+1}{r^2-r+4}\right) &= \frac{r^2-6r-1}{r^2-r+4}\end{aligned}$$

Numerically, this has solutions at 0.250706931 and 0.83089422. The first is a local maximum, and the second is a local minimum. We can quickly check that $r=0.83089422$ gives a lower value than $r=0$, so we aim to achieve $r=0.83089422$.

That is, we are aiming to solve

$$\frac{256a + 63b + 46c}{\sqrt{3538464a + 475131b + 526884c}} = 0.83089422$$

$$256a + 63b + 46c = 0.83089422\sqrt{3538464a + 475131b + 526884c}$$

$$(256a + 63b + 46c)^2 = 0.83089422^2(3538464a + 475131b + 526884c)$$

We have 200 and $b = 96$, so the equation becomes

$(510224230.309 - \sqrt{510224230.309^2 - 4 * 7396 * 2757262617.69})/2/7396$

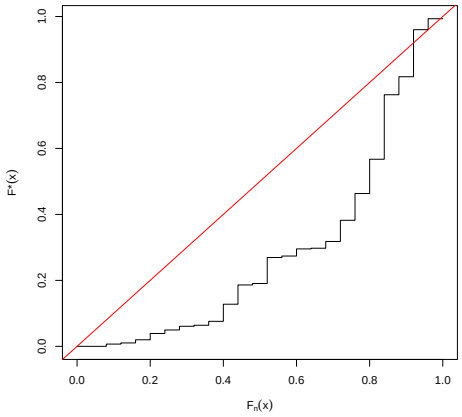
5.40444469984

68981.1058728

$$(57248 + 86c)^2 = 0.83089422^2(753305376 + 477604c)$$

$7396c^2 - 510224230.309c + 2757262617.69 = 0c$ $= 68981.1058728$ or 5.

7. *An insurance company collects a sample of 25 past claims, and attempts to fit a distribution to the claims. Based on experience with other claims, the company believes that an exponential distribution with mean $\theta = 2,400$ 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 3,000?

We have that $F^*(3000) = 1 - e^{-\frac{3000}{2400}} = 0.7134952$, so we see from the plot that the corresponding $F_n(3000) = 0.84$, which corresponds to $0.84 \times 25 = 21$ samples.

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

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

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

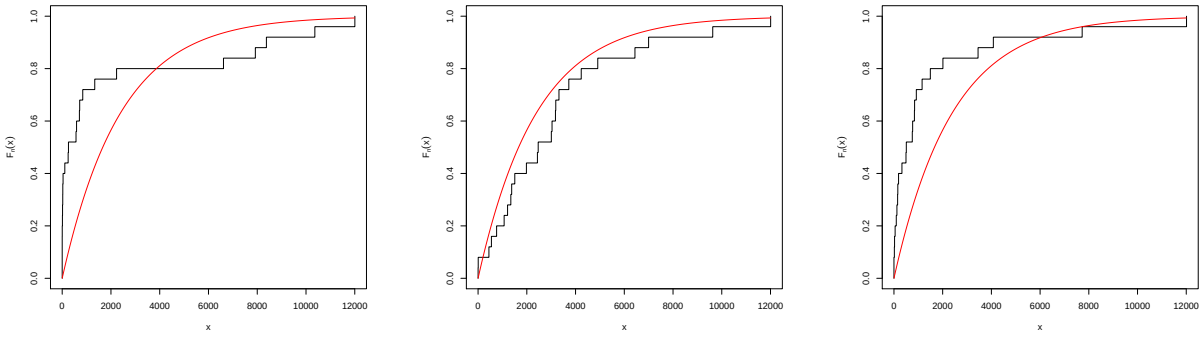
(iii) The exponential distribution assigns too much probability to tail values and too little probability to central values.

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

(i) — From the $p-p$ plot, we see that $F_n(x) > F^*(x)$ for almost all x, which means that the model assigns too little probability to small values of x.

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

(i) (ii) (iii)



From the $p-p$ plot, we have that $F_n(x) < F^*(x)$ for nearly all x. This is the case in plot (iii) but not in the other plots.