ACSC/STAT 4703, Actuarial Models II

Fall 2018

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*
Engineer	*1300*	*0.015*	*\$46,000*	*\$88,000*
Salesperson	*1100*	*0.005*	*\$29,000*	*\$32,000*
Manager	*150*	*0.001*	*\$20,000*	*\$28,000*

Calculate the cost of reinsuring losses above \$3,000,000, 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.

We calculate the expectation and variance of the aggregate losses for each class of workers:

Class	Expected aggregate claims	variance of aggregate claims
Engineer	$1300 \times 0.015 \times 46000 = 897000$	$1300 \times 0.015 \times 88000^2 + 1300 \times 0.015 \times (1 - 0.015) \times 46000^2 = 191651070000$
Salesperson	$1100 \times 0.005 \times 29000 = 159500$	$1100 \times 0.005 \times 32000^2 + 1100 \times 0.005 \times (1 - 0.005) \times 29000^2 = 10234372500$
Manager	$150 \times 0.001 \times 20000 = 3000$	$150 \times 0.001 \times 28000^2 + 150 \times 0.001 \times (1 - 0.001) \times 20000^2 = 177540000$

The expected aggregate loss on the whole portfolio is therefore, $897000 + 159500 + 3000 = \$1,059,500$, and the variance of the aggregate loss is $191651070000 + 10234372500 + 177540000 = 202062982500$. We get the parameters of the gamma approximation by matching moments:

$$\begin{aligned}
\alpha\theta &= 1059500 \\
\alpha\theta^2 &= 202062982500 \\
\alpha &= \frac{1059500^2}{202062982500} \\
&= 5.55539780771 \\
\theta &= \frac{1059500}{5.55539780771} \\
&= 190715.41529
\end{aligned}$$

Now we want to use this approximation to calculate the expectation and variance of the payments on the reinsurance. The expected reinsurance payment is

$$\begin{aligned}
&\int_{3000000}^{\infty}(x-3000000)\frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)}\,dx\\
=&\alpha\theta\int_{3000000}^{\infty}\frac{x^{\alpha}e^{-\frac{x}{\theta}}}{\theta^{\alpha+1}\Gamma(\alpha+1)}\,dx-3000000\int_{3000000}^{\infty}\frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)}\,dx\\
=&\$253.2817
\end{aligned}$$

This can be calculated using the following `R` code.

```
th<-190715.41529
al<-5.55539780771
integral1<-pgamma(3000000,shape=al+1,scale=th,lower.tail=FALSE)
integral2<-pgamma(3000000,shape=al,scale=th,lower.tail=FALSE)
al*th*integral1 -3000000*integral2
```

The expected square of the reinsurance payment is

$$\begin{aligned}
&\int_{3000000}^{\infty}(x-3000000)^2\frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)}\,dx\\
=&\alpha(\alpha+1)\theta^2\int_{3000000}^{\infty}\frac{x^{\alpha+1}e^{-\frac{x}{\theta}}}{\theta^{\alpha+2}\Gamma(\alpha+2)}\,dx-6000000\alpha\theta\int_{3000000}^{\infty}\frac{x^{\alpha}e^{-\frac{x}{\theta}}}{\theta^{\alpha+1}\Gamma(\alpha+1)}\,dx\\
&\qquad+9\times10^{12}\int_{3000000}^{\infty}\frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)}\,dx\\
=&126320327
\end{aligned}$$

This is calculated using the following `R` code.

```
integral1<-pgamma(3000000,shape=al+2,scale=th,lower.tail=FALSE)
integral2<-pgamma(3000000,shape=al+1,scale=th,lower.tail=FALSE)
integral3<-pgamma(3000000,shape=al,scale=th,lower.tail=FALSE)
al*(al+1)*th^2*integral1 -6000000*al*th*integral2 +3000000^2*integral3
```

The standard deviation of the reinsurance payment is therefore $\sqrt{126320327-253.2817^2}=$ \$11236.3773246. The premium is therefore $253.28+11236.38=\$11,489.66$.

2. *An insurance company is modelling claim data as following a Weibull distribution with $\tau=0.7$. It collects the following sample of claims:*

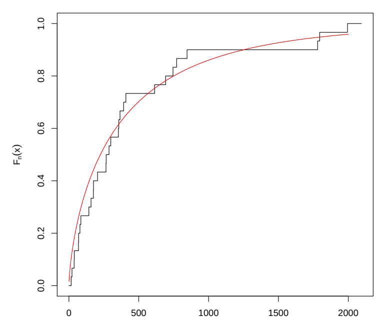
   16.3 22.3 37.5 38.6 68.6 69.7 79.1 85.8 142.9 158.5 175.2 176.1 205.1 265.5 266.9 287.3 299.8 354.2 357.4 365.9 391.9 407.9 613.4 692.4 745.2 771.3 845.9 1780.3 1795.5 1994.7

The MLE for θ is 380.1094. Graphically compare this empirical distribution with the best fitting Weibull distribution with $\tau = 0.7$. Include the following plots:

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

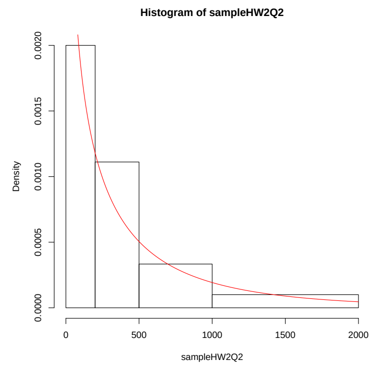
```
FxPlot<-function(x){
    n<-length(x)
    Fx<-seq_len(n)/n
    xv<-sort(x)
    xvals<-c(as.vector(rbind(c(0,xv[seq_len(n-1)]),xv)),xv[n],xv[n]*1.05)
    distvals<-c(0,as.vector(rbind(c(0,Fx[seq_len(n-1)]),Fx)),1)
    plot(xvals,distvals,ylim=c(0,1),xlab=expression(x),ylab=expression(F[n](
}
sampleHW2Q2<-c(
 16.3, 22.3, 37.5, 38.6, 68.6, 69.7, 79.1,  85.8, 142.9, 158.5,
175.2,176.1,205.1,265.5,266.9,287.3,299.8, 354.2, 357.4, 365.9,
391.9,407.9,613.4,692.4,745.2,771.3,845.9,1780.3,1795.5,1994.7)

FxPlot(sampleHW2Q2)
points(1:2000,1-exp(-((1:2000)/380.1094)^0.7),col="red",type='l')
```



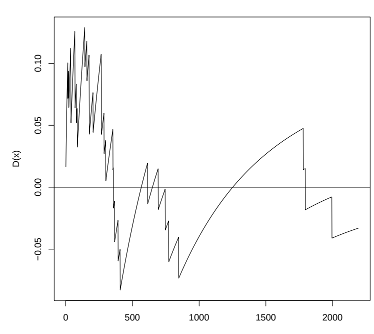
(b) Comparisons of $f(x)$ and $f^*(x)$

```
hist(sampleHW2Q2,breaks=c(0,200,500,1000,2000))
x<-1:2000
fx<-0.7*(x^(-0.3)/380.1094^0.7)*exp(-(x/380.1094)^(0.7))
points(x,fx,type='l',col="red")
```



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

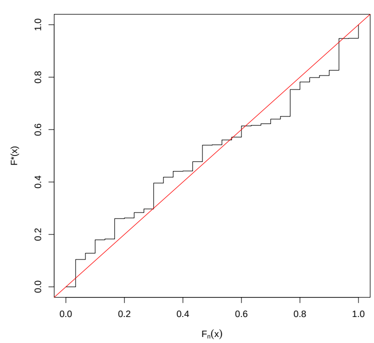
```
DxPlot<-function(x,F){
    n<-length(x)
    plotvals<-(max(x)*1.1)*(1:2000)/2000
    Fstx<-F(plotvals)
    xv<-sort(x)
    Fnx<-rep(0,2000)
    cval<-1
    for(i in 1:2000){
        while(cval<=n&xv[cval]<plotvals[i]){
            cval<-cval+1
        }
        Fnx[i]<-cval-1
    }
    Fnx<-Fnx/n
    plot(plotvals,Fstx-Fnx,xlab=expression(x),ylab="D(x)",type='l')
    abline(h=0)
}
Fweibul<-function(x){1-exp(-(x/380.1094)^0.7)}
DxPlot(sampleHW2Q2,Fweibul)
```



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

```
ppPlot<-function(dist){
    n<-length(dist)
    xv<-seq_len(n)/n
    xvals<-c(as.vector(rbind(c(0,xv[seq_len(n-1)]),xv)),xv[n],1)
    distsort<-sort(dist)
    distvals<-c(0,as.vector(rbind(c(0,distsort[seq_len(n-1)]),distsort)),1)
    plot(xvals,distvals,xlim=c(0,1),ylim=c(0,1),xlab=expression(F[n](x)),yla
    abline(0,1,col="red")
}

ppPlot(Fweibul(sampleHW2Q2))
```



3. *For the data in Question 2, calculate the following test statistics for the goodness of fit of the Weibull distribution with $\tau = 0.7$ and $\theta = 380.1094$:*

   (a) The Kolmogorov-Smirnov test.

   For the observed data points we calculate:

x	$F^*(x)$	$F_n(x^+)$	$F_n(x^-)$	$D(x^+)$	$D(x^+)$
16.3	0.10444	0.033333	0	−0.07111	0.10444
22.3	0.12835	0.066667	0.033333	−0.06168	0.09502
37.5	0.17934	0.1	0.066667	−0.07934	0.11267
38.6	0.18265	0.13333	0.1	−0.04931	0.08265
68.6	0.26040	0.16667	0.13333	−0.09373	0.12706
69.7	0.26289	0.2	0.16667	−0.06289	0.09622
79.1	0.28342	0.23333	0.2	−0.05009	0.08342
85.8	0.29727	0.26667	0.23333	−0.03060	0.06394
142.9	0.39600	0.3	0.26667	−0.09600	**0.12933**
158.5	0.41848	0.33333	0.3	0.08514	0.11848
175.2	0.44093	0.36667	0.33333	−0.07427	0.10760
176.1	0.44210	0.4	0.36667	−0.04210	0.07543
205.1	0.47758	0.43333	0.4	−0.04425	0.07758
265.5	0.54062	0.46667	0.43333	−0.07395	0.10728
266.9	0.54193	0.5	0.46667	−0.04193	0.07527
287.3	0.56047	0.53333	0.5	−0.02714	0.06047
299.8	0.57127	0.56667	0.53333	−0.00460	0.03794
354.2	0.61395	0.6	0.56667	−0.01395	0.04728
357.4	0.61626	0.63333	0.6	0.01707	0.01626
365.9	0.62231	0.66667	0.63333	0.04436	−0.01102
391.9	0.63999	0.7	0.66667	0.06001	−0.02668
407.9	0.65028	0.73333	0.7	0.08305	−0.04972
613.4	0.75289	0.76667	0.73333	0.01378	0.01956
692.4	0.78165	0.8	0.76667	0.01835	0.01498
745.2	0.79850	0.83333	0.8	0.03483	−0.00150
771.3	0.80622	0.86667	0.83333	0.06044	−0.02711
845.9	0.82633	0.9	0.86667	0.07367	−0.04034
1780.3	0.94751	0.93333	0.9	−0.01418	0.04751
1795.5	0.94843	0.96667	0.93333	0.01824	0.01510
1994.7	0.95888	1	0.96667	0.04112	−0.00778

We see that the Kolmogorov-Smirnov statistic is 0.12933.

(b) The Anderson-Darling test.

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

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

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

y_j	$F_n(y_j)$	$F^*(y_j)$	$(1-F_n(y_j))^2$ $(\log(1-F^*(y_j))-\log(1-F^*(y_{j+1})))$	$F_n(y_j)^2$ $(\log(F^*(y_{j+1}))-\log(F^*(y_j)))$
0.0	0	0	0.110305024718	0
16.3	$\frac{1}{30}$	0.10444	0.0252864239085	0.000229048987189
22.3	$\frac{2}{30}$	0.12835	0.0525102345225	0.00148679643667
37.5	$\frac{3}{30}$	0.17934	0.00327294884466	0.0001828478896
38.6	$\frac{4}{30}$	0.18265	0.0750777951422	0.00630478805402
68.6	$\frac{5}{30}$	0.26040	0.002345612875	0.000264773767223
69.7	$\frac{6}{30}$	0.26289	0.0180763512058	0.003007461036
79.1	$\frac{7}{30}$	0.28342	0.0114720071619	0.0025976706374
85.8	$\frac{8}{30}$	0.29727	0.0814196408298	0.0203930456357
142.9	$\frac{9}{30}$	0.39600	0.0185833672835	0.00496889066205
158.5	$\frac{10}{30}$	0.41848	0.0175012417658	0.00580738858833
175.2	$\frac{11}{30}$	0.44093	0.000838059693796	0.000355319113467
176.1	$\frac{12}{30}$	0.44210	0.0236586166866	0.0123531888293
205.1	$\frac{13}{30}$	0.47758	0.0412890988326	0.0232792948074
265.5	$\frac{14}{30}$	0.54062	0.000816066001634	0.000529509110947
266.9	$\frac{15}{30}$	0.54193	0.0103271265508	0.0084080974435
287.3	$\frac{16}{30}$	0.56047	0.00541743564968	0.00542839014342
299.8	$\frac{17}{30}$	0.57127	0.0196895853329	0.0231356317342
354.2	$\frac{18}{30}$	0.61395	0.00096176897168	0.00135408509892
357.4	$\frac{19}{30}$	0.61626	0.00213619629605	0.00391800268022
365.9	$\frac{20}{30}$	0.62231	0.00532560208722	0.0124478374173
391.9	$\frac{21}{30}$	0.63999	0.0026118629379	0.007821583938
407.9	$\frac{22}{30}$	0.65028	0.0246963987151	0.0787906814834
613.4	$\frac{23}{30}$	0.75289	0.00673584568866	0.0220320003214
692.4	$\frac{24}{30}$	0.78165	0.0032130064972	0.0136524751034
745.2	$\frac{25}{30}$	0.79850	0.00108533711945	0.00668272047847
771.3	$\frac{26}{30}$	0.80622	0.0019476831447	0.0185037754384
845.9	$\frac{27}{30}$	0.82633	0.0119659580105	0.110845424592
1780.3	$\frac{28}{30}$	0.94751	0.0000781844932452	0.000841038366546
1795.5	$\frac{29}{30}$	0.94843	0.000251766097184	0.0102456496497
1994.7	$\frac{30}{30}$	0.95888	0	0.0419842986768
			0.578896247069	0.447851716119

The Anderson-Darling statistic is therefore $30(0.578896247069+0.447851716119-1) = 0.8024388957$. For a fully specified distribution, the critical value is 2.492: for the Weibull distribution with one parameter estimated, the critical value is even higher, so we cannot reject the Weibull distribution with $\alpha = 5$.

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

The observed frequencies of these intervals are 12, 9 and 9 respectively. Under the Weibull model, the expected frequencies are $30\left(1-e^{-\left(\frac{200}{380.1094}\right)^{0.7}}\right) = 14.1487266103$, $30\left(e^{-\left(\frac{200}{380.1094}\right)^{0.7}} - e^{-\left(\frac{400}{380.1094}\right)^{0.7}}\right) = 5.20884661662$ and

$30e^{-\left(\frac{400}{380.1094}\right)^{0.7}} = 10.6424267731$ respectively.

The chi-squared statistic is therefore

$$\frac{(12-14.1487266103)^2}{14.1487266103}+\frac{(9-5.20884661662)^2}{5.20884661662}+\frac{(9-10.6424267731)^2}{10.6424267731}=3.3391078709$$

There are 3 classes, which gives 2 degrees of freedom, and one estimated parameter reduces this to 1 degree of freedom. For a chi-squared distribution with one degree of freedom, at the 5% significance level, the critical value is 3.841459, so we cannot reject the Weibull distribution.

4. *For the data in Question 2, perform a likelihood ratio test to determine whether a Weibull distribution with fixed $\tau = 0.7$, or a Weibull distribution with τ freely estimated is a better fit for the data. [The MLE for the general Weibull distribution is $\tau = 0.9089666$ and $\theta = 428.7284682$.]*

[The original version on the homework mistakenly gave the MLE for the general Weibull as $\tau = 0.3125$ and $\theta = 295.7674$. This leads to a log-likelihood that is smaller than the log-likelihood for $\theta = 0.7$, at which point it should be obvious that this is not the MLE. This model solution uses the correct MLE.]

The log-likelihood of the Weibull distribution is

$$\sum \log\left(\tau\frac{{x_i}^{\tau-1}}{\theta^\tau}e^{-\left(\frac{x_i}{\theta}\right)^\tau}\right)=n\log\tau+(\tau-1)\sum\log(x_i)-n\tau\log(\theta)-\sum\left(\frac{x_i}{\theta}\right)^\tau$$

We get

$$\frac{dl}{d\theta}=\frac{n\tau}{\theta}-\tau\frac{\sum {x_i}^\tau}{\theta^{\tau+1}}$$

so the MLE for θ is

$$\left(\frac{\sum {x_i}^\tau}{n}\right)^{\frac{1}{\tau}}$$

Substituting this into the above expression for the log-likelihood gives us

$$\begin{aligned}
l(\tau) &= n\log\tau+(\tau-1)\sum\log(x_i)-n\log\left(\frac{\sum {x_i}^\tau}{n}\right)-n\\
\frac{dl(\tau)}{d\tau} &= \frac{n}{\tau}+\sum\log(x_i)-n\frac{\sum {x_i}^\tau\log(x_i)}{\sum {x_i}^\tau}\\
\frac{d^2l(\tau)}{d\tau^2} &= -\frac{n}{\tau^2}-n\frac{\sum {x_i}^\tau\log(x_i)^2}{\sum {x_i}^\tau}+n\left(\frac{\sum {x_i}^\tau\log(x_i)}{\sum {x_i}^\tau}\right)^2
\end{aligned}$$

for $\tau = 0.7, \theta = 380.1094$ $l(\tau)$ is -214.6077, while for $\tau = 0.9089666, \theta = 428.7284682$, it is -213.0527. The log-likelihood ratio statistic is therefore

$2(-213.0527-(-214.6077)) = 3.11$. The null distribution is chi-squared with one degree of freedom, so the critical value at the 5% significance level is 3.841459, so there is not strong evidence against $\tau = 0.7$.

5. *For the data in Question 2, use AIC and BIC to choose between a Weibull distribution with $\tau = 0.7$ and a Pareto distribution for the data. [The MLE for the Pareto distribution is $\alpha = 4.8761$ and $\theta = 1760.6118$.]*

   From Question 4, we have that the log-likelihood for the Weibull distribution is -214.6077. The log-likelihood for the Pareto distribution is

   $$\sum \log\left(\frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}\right) = n\log(\alpha)+n\alpha\log(\theta)-(\alpha+1)\log(x+\theta) = -212.8249$$

   Therefore the AIC is $-214.6077-2 = -216.6077$ for the Weibull distribution and $-212.8249-4 = -216.8249$ for the Pareto distribution. Therefore, the Weibull distribution is prefered.

   The BIC for the Weibull distribution is $-214.6077-\frac{\log(30)}{2} = -216.308298691$ and for the Pareto distribution, it is $-212.8249-2\frac{\log(30)}{2} = -216.226097382$, so under BIC, the Pareto distribution is prefered.

Standard Questions

6. *An insurance company insures three types of properties and has the following estimates:*

Property type	*Probability of claim*	*mean claim*	*standard deviation*
Residential (House)	*0.004*	*\$8,600*	*\$25,800*
Residential (Apartment)	*0.009*	*\$2,300*	*\$6,900*
Commercial	*0.02*	*\$3,600*	*\$12,400*

   The insurance company estimates the mean μ and standard deviation σ for the aggregate loss distribution, and buys stop-loss insurance for losses above \$200,000. One reinsurer models aggregate losses as following a Pareto distribution and sets its premium as 110% of the expected claims on the stop-loss policy. Another reinsurer models aggregate losses as following a Gamma distribution, and sets its premium at 200% of the expected claims. The portfolio includes 2,243 houses and 1,832 apartments. How many commercial properties would it need to include for the two reinsurance companies to charge the same premium on the stop-loss insurance?

   (i) 640

   (ii) 1,209

   (iii) 1,853

   (iv) 2,177

We calculate the expectation and variance of the loss for a property of each type:

Policy	Expected aggregate claims	variance of aggregate claims
Residential (House)	$0.004 \times 8600 = 34.4$	$0.004 \times 25800^2 + 0.004 \times 0.996 \times 8600^2 = 2957216.64$
Residential (Apartment)	$0.009 \times 2300 = 20.7$	$0.009 \times 6900^2 + 0.009 \times 0.991 \times 2300^2 = 475671.51$
Commercial	$0.02 \times 3600 = 72$	$0.02 \times 12400^2 + 0.02 \times 0.98 \times 3600^2 = 3329216$

If the portfolio includes C commercial properties, then the overall aggregate loss has expectation $2243 \times 34.4 + 1832 \times 20.7 + 72C = 72C + 115081.6$, and variance $2243 \times 2957216.64 + 1832 \times 475671.51 + 3329216C = 3329216C + 7504467129.84$

For the first reinsurer, using the Pareto distribution to model aggregate losses, the parameters are obtained by solving

$$\begin{aligned}
\frac{\theta}{\alpha-1} &= 72C + 115081.6 \\
\frac{\theta^2}{(\alpha-1)^2(\alpha-2)} &= 3329216C + 3237903929.84 \\
\alpha - 2 &= \frac{(72C + 115081.6)^2}{3329216C + 3237903929.84} \\
\theta &= (72C + 115081.6)\left(\frac{(72C + 115081.6)^2}{3329216C + 3237903929.84} + 1\right)
\end{aligned}$$

For a Pareto distribution, the expected payment on the excess-of-loss insurance is

$$\begin{aligned}
\int_{1000000}^{\infty} \left(\frac{\theta}{\theta+x}\right)^{\alpha} dx &= \theta \int_{1+\frac{1000000}{\theta}}^{\infty} u^{-\alpha}\, du \\
&= \theta \left[-\frac{u^{1-\alpha}}{\alpha-1}\right]_{1+\frac{1000000}{\theta}}^{\infty} \\
&= \frac{\theta}{\alpha-1}\left(1 + \frac{1000000}{\theta}\right)^{1-\alpha} \\
&= \frac{\theta}{\alpha-1}\left(\frac{\theta}{\theta+1000000}\right)^{\alpha-1}
\end{aligned}$$

For the second reinsurer, the parameters of the Gamma distribution are obtained by solving

$$\begin{aligned}
\alpha\theta &= 72C + 115081.6 \\
\alpha\theta^2 &= 3329216C + 3237903929.84 \\
\alpha &= \frac{(72C + 115081.6)^2}{3329216C + 3237903929.84} \\
\theta &= \frac{3329216C + 3237903929.84}{72C + 115081.6}
\end{aligned}$$

For the Gamma distribution, the expected payment on the excess-of-loss reinsurance is

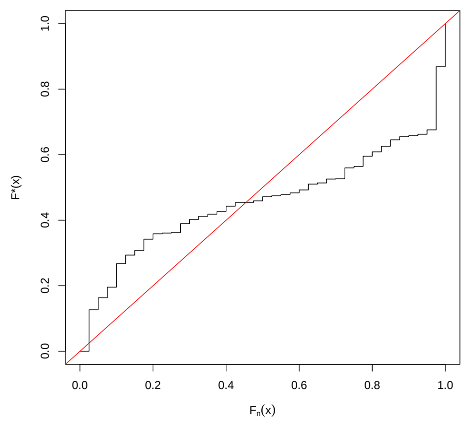
$$\begin{aligned}
&\int_{2000000}^{\infty} (x - 2000000) \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha \Gamma(\alpha)}\,dx \\
=&\alpha\theta \int_{2000000}^{\infty} \frac{x^{\alpha}e^{-\frac{x}{\theta}}}{\theta^{\alpha+1} \Gamma(\alpha+1)}\,dx - 2000000 \int_{2000000}^{\infty} \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha \Gamma(\alpha)}\,dx
\end{aligned}$$

We compute the reinsurance costs for different values of C:

C	Expected Agg. Loss	Variance Agg. Loss	1st Reinsurance Premium	2nd Reinsurance Premium
640	161161.6	9635165370	85562.75	48856.98
1,209	202129.6	11529489274	117262.24	85532.19
1,853	248497.6	13673504378	155960.73	142331.32
2,177	271825.6	14752170362	176328.86	176337.65

So they would need to sell (iv) 2177 commercial policies for the reinsurance to have the same costs.

7. *An insurance company collects a sample of 40 past claims, and attempts to fit a distribution to the claims. Based on experience with other claims, the company believes that a Pareto distribution with $\alpha = 3$ and $\theta = 1,200$ 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 168?

For the Pareto distribution, $F^*(168) = 1 - \left(\frac{1200}{1200+168}\right)^3 = 0.325028483798$. From the p-p plot, we get that $F_n(x) = 0.15$, so there are 6 points less than 168.

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

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

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

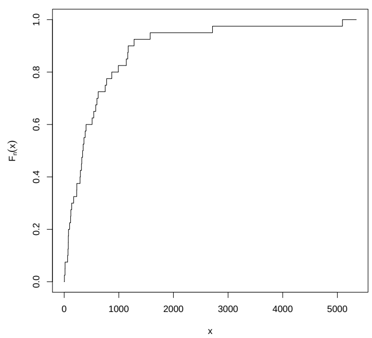
(iii) The Pareto distribution assigns too much probability to tail values and too little probability to central values.

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

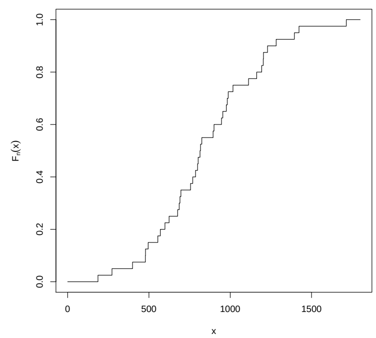
We see that for smaller values, the plot is above the $y = x$ line, meaning $F^*(x) > F_n(x)$, while for larger values, the plot is below the line, meaning $F^*(x) < F_n(x)$. This means that the Pareto distribution assigns too much probability to tail values, so (iii) is the best description.

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

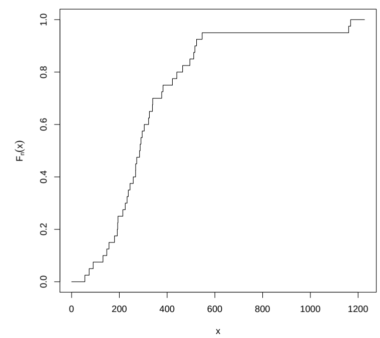
(i)



(ii)



(iii)



Looking at the value $F_n(900)$, we can see $F^*(900) = 1 - \left(\frac{1200}{1200+900}\right)^3 = 0.813411078717$. From the p-p plot, we can read the corresponding value $F_n(900) = 0.975$. For plots (i) and (ii) $F_n(900)$ is clearly less than 0.975, so (iii) is the only possible plot for the empirical distribution.