ACSC/STAT 4703, Actuarial Models II
Fall 2020
Toby Kenney
Homework Sheet 2
Model Solutions

Basic Questions

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

Type of driver	Number	Probability of claim	mean claim	standard deviation
Learner	1100	0.035	$3,900	$34,800
Normal	700	0.015	$3,400	$31,400
Advanced	400	0.011	$2,800	$32,100

Calculate the cost of reinsuring losses above $1,000,000, if the loading on the reinsurance premium is one standard deviation above the expected claim payment on the reinsurance policy using a Pareto approximation for the aggregate losses on this portfolio.

We calculate the expectation and variance for aggregate losses in each class

Type of driver	Expected no. of claims	Variance of no. of claims	mean aggregate loss	variance of aggregate loss
Learner	38.5	37.1525	150150	47190129525
Normal	10.5	10.3425	35700	10472139300
Advanced	4.4	4.3516	12320	4567920544

Therefore the expectation and variance of the aggregate loss are $150150 + 35700 + 12320 = 198170$ and $47190129525 + 10472139300 + 4567920544 = 62230189369$ respectively. For a Pareto distribution, we calculate the moment estimates

$$\frac{\theta}{\alpha - 1} = 198170$$

$$\frac{\theta^2 \alpha}{(\alpha - 1)^2(\alpha - 2)} = 1877899049.48$$

$$\frac{\alpha}{\alpha - 2} = \frac{62230189369}{198170^2} = 1.58462062323$$

$$\alpha = \frac{2 \times 1.58462062323}{0.58462062323} = 5.42102197653$$

$$\theta = 198170 \times 4.42102197653 = 876113.925089$$

Using this distribution, the expected payment on the stop loss insurance is

$$\mathbb{E}((X - 1000000)_+) = \int_{1000000}^{\infty} \frac{\theta^\alpha}{(\theta + x)^\alpha} \, dx$$

$$= \theta^\alpha \int_{1000000 + \theta}^{\infty} u^{-\alpha} \, du$$

$$= \theta^\alpha \left[-\frac{u^{1-\alpha}}{\alpha - 1} \right]_{1000000 + \theta}^{\infty}$$

$$= \frac{\theta^\alpha}{(\alpha - 1)(1000000 + \theta)^{\alpha - 1}}$$

$$= \frac{876113.925089^{5.42102197653}}{(4.42102197653)(1000000 + 876113.925089)^{4.42102197653}}$$

$$= 6839.30585606$$

The expected square of the payment on the stop loss insurance is

$$\mathbb{E}((X - 1000000)_+^2) = 2 \int_{1000000}^{\infty} (x - 1000000) \frac{\theta^\alpha}{(\theta + x)^\alpha} \, dx$$

$$= 2\theta^\alpha \int_{1000000 + \theta}^{\infty} (u - 1000000 - \theta) u^{-\alpha} \, du$$

$$= 2\theta^\alpha \int_{1000000 + \theta}^{\infty} u^{1-\alpha} - (1000000 + \theta) u^{-\alpha} \, du$$

$$= 2\theta^\alpha \left[(1000000 + \theta) \frac{u^{1-\alpha}}{\alpha - 1} - \frac{u^{2-\alpha}}{\alpha - 2} \right]_{1000000 + \theta}^{\infty}$$

$$= \frac{2\theta^\alpha}{(1000000 + \theta)^{\alpha - 2}} \left(\frac{1}{\alpha - 2} - \frac{1}{\alpha - 1} \right)$$

$$= \frac{2 \times 876113.925089^{5.42102197653}}{4.42102197653 \times 3.42102197653 \times (1000000 + 876113.925089)^{3.42102197653}}$$

$$= 7501452514.79$$

so the standard deviation is $\sqrt{7501452514.79 - 6839.30585606^2} = 86340.4679753$. Thus the premium on the stop-loss insurance is $6839.30585606 + 86340.4679753 = \$93,179.77$.

2. *An insurance company is modelling claim data as following a Pareto distribution with $\alpha = 2.6$. It collects the following sample of claims:*

    0.1 0.3 0.7 1.5 4.7 6.7 6.9 7.4 7.8 14.0 20.6 20.8
    22.1 24.3 38.1 44.7 70.7 157.0 244.1 254.6 280.1 282.0
   *285.3 424.8 928.3 1119.9 1694.3 2792.2 2979.6 3613.8*

The MLE for θ is 405.5201. Graphically compare this empirical distribution with the best fitting Pareto distribution with $\alpha = 2.6$. 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(
    0.1,    0.3,    0.7,    1.5,    4.7,     6.7,     6.9,     7.4,
7.8,    14.0,   20.6,   20.8,   22.1,   24.3,   38.1,    44.7,    70.7,   157.0,   244.1,   254.6,

FxPlot(sampleHW2Q2)
points(1:4000,1-405.5201^2.6/(405.5201+(1:4000))^2.6,col="red",type='l')
```
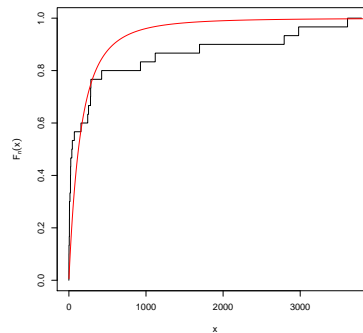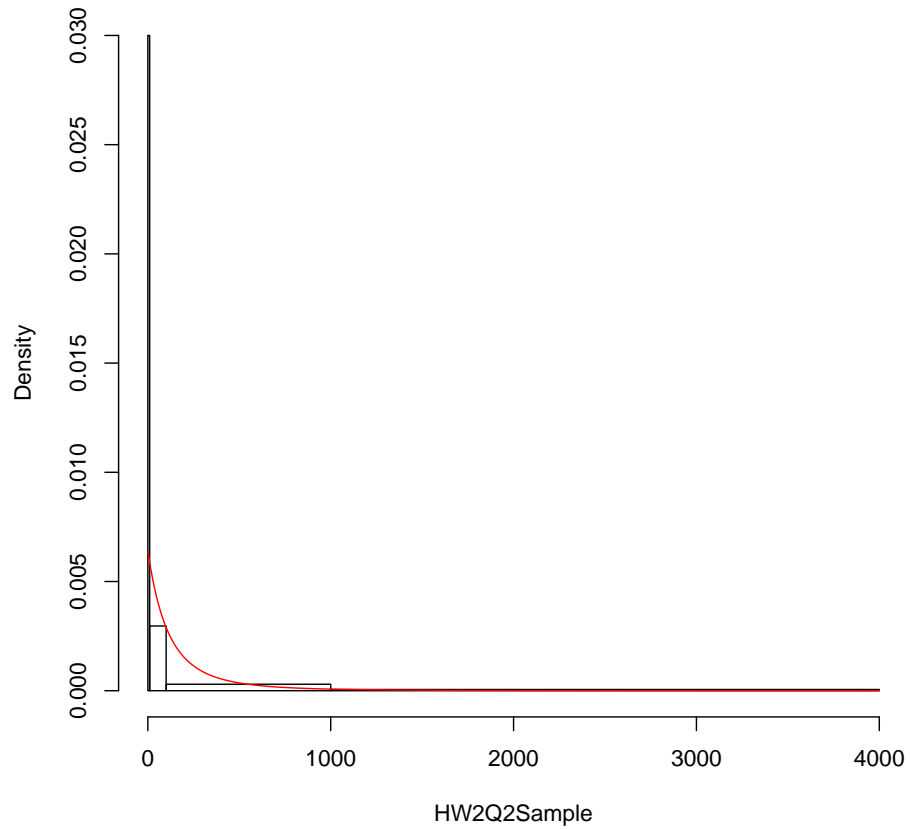

(b) Comparisons of $f(x)$ and $f^*(x)$

```
hist(HW2Q2Sample,breaks=c(0,10,100,1000,4000))
points(1:4000,2.6*405.5201^2.6/(405.5201+(1:4000))^3.6,type='l',col="red")
```

Histogram of HW2Q2Sample

(c) A plot of D(x) against x.
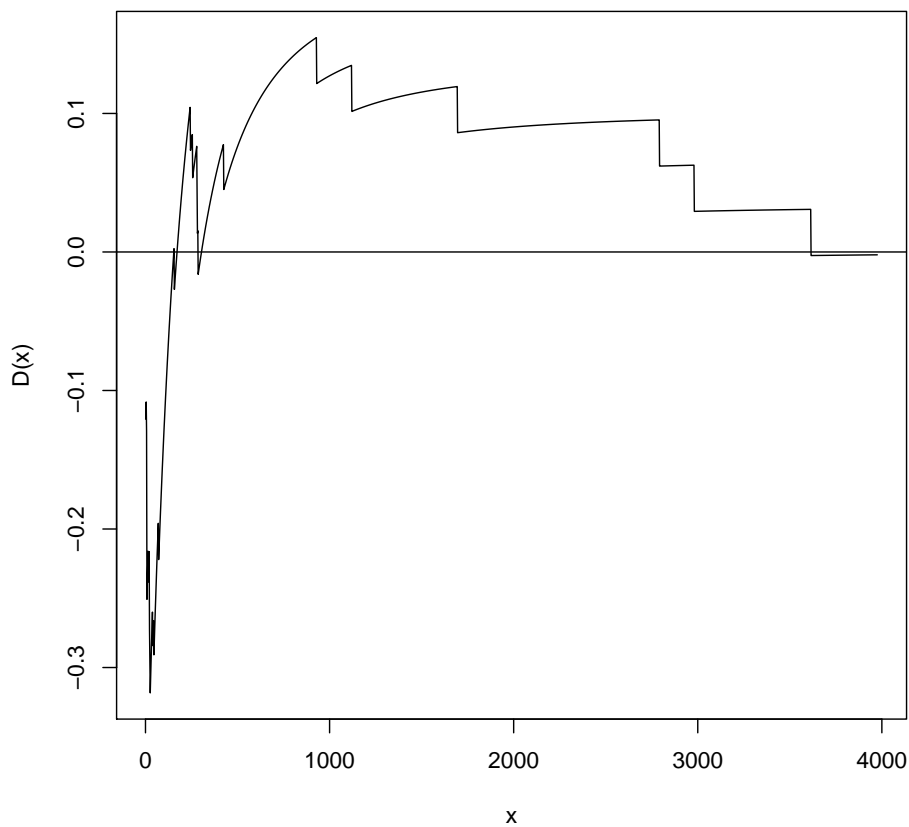
```
DxPlot<-function(x,F){
    n<-length(x)
    plotvals<-(max(x)*1.1)*(1:2000)/2000
    Fstx<-F(plotvals)
    xv<-sort(x)
    Fnx<-rep(0,2000)
    cval<-1
    for(i in 1:2000){
        while(cval<=n&xv[cval]<plotvals[i]){
            cval<-cval+1
        }
        Fnx[i]<-cval-1
    }
    Fnx<-Fnx/n
```

```
        plot(plotvals,Fstx−Fnx,xlab=expression(x),ylab="D(x)",type='l')
        abline(h=0)
}

Fpareto<−function(x){return(1−(405.5201/(405.5201+x))^2.6)}
DxPlot(HW2Q2Sample,Fpareto)
```


(d) A p-p plot of $F(x)$ against $F^*(x)$.
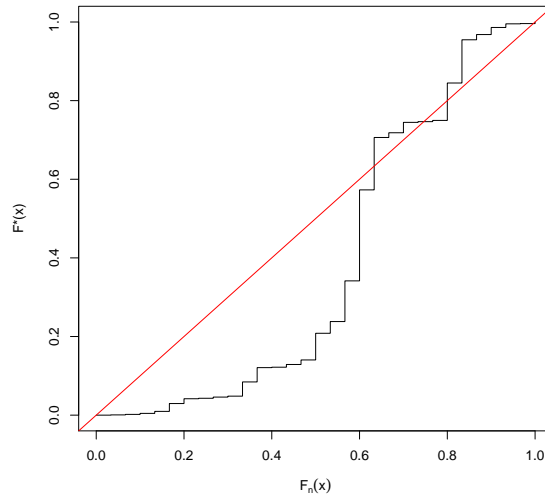
```
ppPlot<−function(dist){
    n<−length(dist)
    xv<−seq_len(n)/n
    xvals<−c(as.vector(rbind(c(0,xv[seq_len(n−1)]),xv)),xv[n],1)
    distsort<−sort(dist)
    distvals<−c(0,as.vector(rbind(c(0,distsort[seq_len(n−1)]),distsort)),1)
    plot(xvals,distvals,xlim=c(0,1),ylim=c(0,1),xlab=expression(F[n](x)),yla
```

```
        abline(0,1,col="red")
}
```

$$ppPlot(1-(405.5201/(405.5201+HW2Q2Sample))\char`^2.6)$$

3. *For the data in Question 2, calculate the following test statistics for the goodness of fit of the Pareto distribution with $\alpha = 2.6$ and $\theta = 405.5201$:*

   (a) The Kolmogorov-Smirnov test.

   For the observed data points we calculate:

x	$F^*(x)$	$F_n(x^+)$	$F_n(x^-)$	$D(x^+)$	$D(x^-)$
0.1	0.0006408675	0.033333	0	0.0326921325	0.0006408675
0.3	0.0019208974	0.066667	0.033333	0.0647461026	−0.0314121026
0.7	0.0044741555	0.1	0.066667	0.0955258445	−0.0621928445
1.5	0.0095536075	0.13333	0.1	0.1237763925	−0.0904463925
4.7	0.0295164736	0.16667	0.13333	0.1371535264	−0.1038135264
6.7	0.0417112833	0.2	0.16667	0.1582887167	−0.1249587167
6.9	0.0429190732	0.23333	0.2	0.1904109268	−0.1570809268
7.4	0.0459293414	0.26667	0.23333	0.2207406586	−0.1874006586
7.8	0.0483281249	0.3	0.26667	0.2516718751	−0.2183418751
14.0	0.0844649177	0.33333	0.3	0.2488650823	−0.2155350823
20.6	0.1208784034	0.36667	0.33333	0.2457915966	−0.2124515966
20.8	0.1219503014	0.4	0.36667	0.2780496986	−0.2447196986
22.1	0.1288737236	0.43333	0.4	0.3044562764	−0.2711262764
24.3	0.1404191580	0.46667	0.43333	**0.326250842**	−0.292910842
38.1	0.2082227086	0.5	0.46667	0.2917772914	−0.2584472914
44.7	0.2380481751	0.53333	0.5	0.2952818249	−0.2619518249
70.7	0.3415358650	0.56667	0.53333	0.225134135	−0.191794135
157.0	0.5729553513	0.6	0.56667	0.0270446487	0.0062853513
244.1	0.7062907520	0.63333	0.6	−0.072960752	0.106290752
254.6	0.7182833508	0.66667	0.63333	−0.0516133508	0.0849533508
280.1	0.7447210558	0.7	0.66667	−0.0447210558	0.0780510558
282.0	0.7465512445	0.73333	0.7	−0.0132212445	0.0465512445
285.3	0.7496870652	0.76667	0.73333	0.0169829348	0.0163570652
424.8	0.8448349990	0.8	0.76667	−0.044834999	0.078164999
928.3	0.9547540895	0.83333	0.8	−0.1214240895	0.1547540895
1119.9	0.9680831815	0.86667	0.83333	−0.1014131815	0.1347531815
1694.3	0.9860953913	0.9	0.86667	−0.0860953913	0.1194253913
2792.2	0.9953414899	0.93333	0.9	−0.0620114899	0.0953414899
2979.6	0.9959826507	0.96667	0.93333	−0.0293126507	0.0626526507
3613.8	0.9974293944	1	0.96667	0.0025706056	0.0307593944

We see that the Kolmogorov-Smirnov statistic is 0.326250842.

(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	$F_n(y_j)$	$F^*(y_j)$	$(1 - F_n(y_j))^2$ $(\log(1 - F^*(y_j)) - \log(1 - F^*(y_{j+1})))$	$F_n(y_j$ $(\log(F^*(y_{j+1})) - \log(F^*(y_j)$
0.0	0	0	0.000641072943356	
0.1	$\frac{1}{30}$	0.0006408675	0.00119765104272	0.0012196944715
0.3	$\frac{2}{30}$	0.0019208974	0.00223130738088	0.0037578895384
0.7	$\frac{3}{30}$	0.0044741555	0.00414342659473	0.0075860121300
1.5	$\frac{4}{30}$	0.0095536075	0.0152936098555	0.02005386138
4.7	$\frac{5}{30}$	0.0295164736	0.00878147300638	0.0096061983427
6.7	$\frac{6}{30}$	0.0417112833	0.00080713990752	0.0011417859560
6.9	$\frac{7}{30}$	0.0429190732	0.00185162729098	0.0036906710253
7.4	$\frac{8}{30}$	0.0459293414	0.00135381681259	0.0036202331185
7.8	$\frac{9}{30}$	0.0483281249	0.0189686950814	0.0502490327170
14.0	$\frac{10}{30}$	0.0844649177	0.0180379824358	0.0398276485255
20.6	$\frac{11}{30}$	0.1208784034	0.000489366349455	0.0011869409533
20.8	$\frac{12}{30}$	0.1219503014	0.00284985049392	0.0088351107130
22.1	$\frac{13}{30}$	0.1288737236	0.00428428614481	0.0161111262420
24.3	$\frac{14}{30}$	0.1404191580	0.0233712989485	0.0857992800730
38.1	$\frac{15}{30}$	0.2082227086	0.009599205693	0.0334662120620
44.7	$\frac{16}{30}$	0.2380481751	0.0317897358289	0.102678649140
70.7	$\frac{17}{30}$	0.3415358650	0.0813118119102	0.166128470120
157.0	$\frac{18}{30}$	0.5729553513	0.0598877193344	0.075318908402
244.1	$\frac{19}{30}$	0.7062907520	0.00560479392964	0.0067535669228
254.6	$\frac{20}{30}$	0.7182833508	0.0109494366244	0.0160647094770
280.1	$\frac{21}{30}$	0.7447210558	0.0006475672296	0.00120272188140
282.0	$\frac{22}{30}$	0.7465512445	0.000885317715911	0.00225415583290
285.3	$\frac{23}{30}$	0.7496870652	0.0260365747205	0.070230902725
424.8	$\frac{24}{30}$	0.8448349990	0.0492950713172	0.078279979893
928.3	$\frac{25}{30}$	0.9547540895	0.00969386649473	0.0096279204649
1119.9	$\frac{26}{30}$	0.9680831815	0.0147717822263	0.013846794005
1694.3	$\frac{27}{30}$	0.9860953913	0.0109352467035	0.0075595587199
2792.2	$\frac{28}{30}$	0.9953414899	0.000658103861999	0.00056095570259
2979.6	$\frac{29}{30}$	0.9959826507	0.000496089770955	0.0013563697015
3613.8	$\frac{30}{30}$	0.9974293944	0	0.0025739152797
			0.416864927651	0.840589275533

The Anderson-Darling statistic is therefore $30(0.416864927651 + 0.840589275536 - 1) = 7.7236260957$. For a fully specified distribution, the critical value is 2.492 so we reject the Pareto distribution with $\alpha = 2.6$ and $\theta = 405.5201$.

(c) The chi-square test, dividing into the intervals 0–100, 100–500, and more than 500.

The observed frequencies of these intervals are 17, 7 and 6 respectively.

Under the Pareto model, the expected frequencies are $30 \left(1 - \left(\frac{405.5201}{405.5201 + 100} \right)^{2.6} \right) = 13.086497847$, $30 \left(\left(\frac{405.5201}{405.5201 + 100} \right)^{2.6} - \left(\frac{405.5201}{405.5201 + 500} \right)^{2.6} \right) = 13.197992364$

and $30 \left(\frac{405.5201}{405.5201 + 500} \right)^{2.6} = 3.71550978906$ respectively.

The chi-squared statistic is therefore

$$\frac{(17 - 13.086497847)^2}{13.086497847} + \frac{(7 - 13.197992364)^2}{13.197992364} + \frac{(6 - 3.71550978906)^2}{3.71550978906} = 5.48563092712$$

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 reject the Pareto distribution.

4. *For the data in Question 2, perform a likelihood ratio test to determine whether a Pareto distribution with fixed $\alpha = 2.6$, or a Pareto distribution with α freely estimated is a better fit for the data. [The MLE for the general Pareto distribution is $\alpha = 0.4254277$ and $\theta = 9.096$.]*

   For the Pareto distribution with $\alpha = 2.6$, the log-likelihood is

$$\sum_{i=1}^{30} \log(\alpha) + \alpha \log(\theta) - (\alpha + 1) \log(\theta + x_i) = -205.1907$$

   while for the general Pareto distribution with $\alpha = 0.4254277$ and $\theta = 9.096$, the log-likelihood is -192.3921. The log-likelihood ratio is therefore

$$2(-192.3921 - (-205.1907)) = 25.5972$$

   Under the null hypothesis, the log-likelihood ratio should follow a chi-square distribution with 1 degree of freedom, so the critical value is 3.841459. We therefore find that the general Pareto distribution is a significantly better fit for the data than the Pareto with $\alpha = 2.6$.

5. *For the data in Question 2, use AIC and BIC to choose between a Pareto distribution with $\alpha = 2.6$ for the data and a gamma distribution. [The MLE for the gamma distribution is $\alpha = 0.2962591$ and $\theta = 1726.7992261$.]*

   The log-likelihood for the gamma distribution is

$$\sum_{i=1}^{30} (\alpha - 1) \log(x_i) - \alpha \log(\theta) - \log(\gamma(\alpha)) = -190.8537$$

   The gamma distribution has two estimated parameters, while the Pareto distribution has 1. We have that the AIC is given by

$$-190.8537 - 2 \times 2 = -194.8537$$

   for the gamma distribution and

$$-205.1907 - 1 \times 2 = -207.1907$$

for the Pareto, and the BIC is given by

$$-190.8537 - 2\frac{\log(30)}{2} = -194.254897382$$

for the gamma distribution and

$$-205.1907 - \frac{\log(30)}{2} = -206.891298691$$

for the Pareto. Thus, both AIC and BIC prefer the Gamma distribution.

Standard Questions

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

Property type	Probability of claim	mean claim	standard deviation
Train	0.01	$35,600	$594,800
Ship	0.03	$21,300	$334,900
Truck	0.13	$ 8,600	$217,300

The insurance company estimates the mean μ and standard deviation σ for the aggregate loss distribution, and buys stop-loss insurance for losses above $2,500,000. One reinsurer models aggregate losses as following a Pareto distribution and sets its premium as 120% of the expected claims on the stop-loss policy. Another reinsurer models aggregate losses as following a Gamma distribution, and sets its premium at 150% of the expected claims. The company insures 134 trains, 211 ships and 403 trucks. Which reinsurance company is cheaper?

We calculate the expectation and variance for aggregate losses from each line of insurance:

Type of policy	Expected no. of claims	Variance of no. of claims	mean aggregate loss	variance of aggregate loss
Train	1.34	1.3266	47704	475755913376
Ship	6.33	6.1401	134829	712745905269
Truck	52.39	45.5793	450554	2477189648130

The aggregate loss over all lines therefore has expectation $47704+134829+450554 = 633087$ and variance $475755913376+712745905269+2477189648130 = 3665691466775$

Matching moments gives the following parameters for the Pareto distribution

$$\frac{\theta}{\alpha - 1} = 633087$$

$$\frac{\theta^2 \alpha}{(\alpha - 1)^2 (\alpha - 2)} = 3665691466775$$

$$\frac{\alpha}{\alpha - 2} = \frac{3665691466775}{633087^2} = 9.14595619956$$

$$\alpha = \frac{2}{\left(1 - \frac{1}{9.14595619956}\right)} = 2.24552059341$$

$$\theta = 633087 \times 1.24552059341 = 788522.89592$$

Therefore the expected claims on the stop loss insurance policy are given by

$$\mathbb{E}((X - 2500000)_+) = \int_{2500000}^{\infty} \frac{\theta^\alpha}{(\theta + x)^\alpha}\, dx$$

$$= \theta^\alpha \int_{2500000+\theta}^{\infty} u^{-\alpha}\, du$$

$$= \theta^\alpha \left[-\frac{u^{1-\alpha}}{\alpha - 1}\right]_{2500000+\theta}^{\infty}$$

$$= \frac{\theta^\alpha}{(\alpha - 1)(2500000 + \theta)^{\alpha - 1}}$$

$$= \frac{788522.89592^{2.24552059341}}{(1.24552059341)(2500000 + 788522.89592)^{1.24552059341}}$$

$$= 106907.521541$$

so the premium is $1.2 \times 106907.521541 = \$128,289.03$. For the gamma distribution, the parameters are

$$\alpha\theta = 633087$$

$$\alpha\theta^2 = 3665691466775$$

$$\alpha = \frac{633087^2}{3665691466775} = 0.109337938886$$

$$\theta = \frac{633087}{0.109337938886} = 5790185.9725$$

so the estimated payments on the stop-loss insurance are

$$\mathbb{E}((X - 2500000)_+) = \int_{2500000}^{\infty} \frac{(x - 2500000)x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^{\alpha}\Gamma(\alpha)}\,dx$$

$$= \frac{\theta^{-\alpha}}{\Gamma(\alpha)}\left(\int_{2500000}^{\infty} x^{\alpha}e^{-\frac{x}{\theta}}\,dx - 2500000\int_{2500000}^{\infty} x^{\alpha-1}e^{-\frac{x}{\theta}}\,dx\right)$$

$$= \alpha\theta S_{\gamma(\alpha+1)}\left(\frac{2500000}{\theta}\right) - 2500000 S_{\gamma(\alpha)}\left(\frac{2500000}{\theta}\right)$$

$$= 257444.4$$

where $S_{\gamma(\alpha)}$ is the survival function of a gamma distribution with shape parameter α and scale parameter 1. The premium from the company that uses a gamma distribution to model aggregate losses is therefore $257444.4 \times 1.5 = 386166.6$. Therefore, the reinsurance company that uses the Pareto distribution is cheaper.

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 Gamma distribution with $\alpha = 2$ and $\theta = 1,400$ may be appropriate to model these claims. It constructs the following p-p plot to compare the sample to this distribution:*
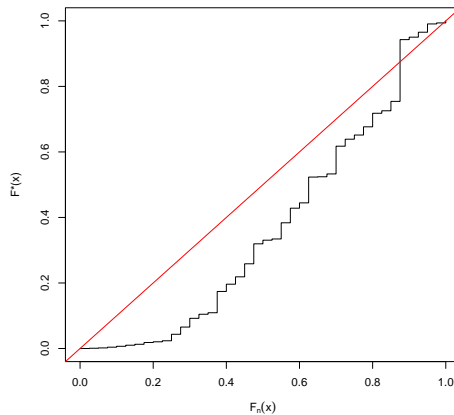
(a) How many of the points in their sample were less than 2,300?

We calculate that $F^*(2300) = 1400F_{\gamma(2)}\left(\frac{2300}{1400}\right) = 0.4888011$. From the graph, we therefore read that $F_n(2300) = 0.625 = \frac{25}{40}$, so there are 25 samples less that 2,300.

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

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

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

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

(iv) The Gamma distribution assigns too much probability to central values and too little probability to tail values.
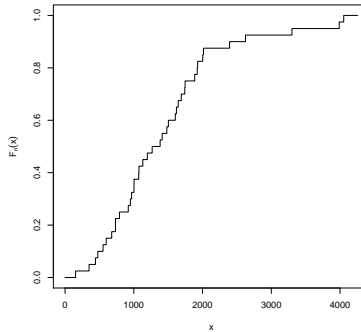
We see that the p-p plot is below the line $y = x$ for small values of $F_n(x)$. This means that for small values x, we have $F_n(x) > F^*(x)$, so (i) best describes the fit.

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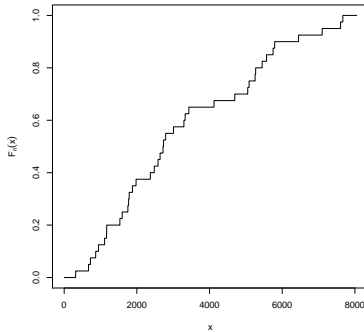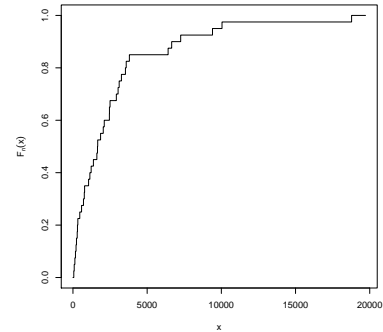
(i)   (ii)   (iii)

We know that $F_n(2,300) = 0.625$, which is clearly not the case in (a) or (b), so (c) must be the empirical distribution.