ACSC/STAT 4703, Actuarial Models II

FALL 2021
Toby Kenney

Homework Sheet 2

Model Solutions

Basic Questions

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

Type of worker	*Number*	*Probability of claim*	*mean claim (millions)*	*standard deviation (millions)*
Family Doctor	*1,500*	*0.00015*	*\$1.3*	*\$0.9*
Surgeon	*600*	*0.00074*	*\$4.4*	*\$5.6*
Nurse	*2,400*	*0.00136*	*\$0.5*	*\$0.4*

They model aggregate losses using a gamma distribution. 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.

We calculate the mean and variance of the aggregate loss:

Type of worker	$\mathbb{E}(N)$	$\mathrm{Var}(N)$ of claim	mean aggregate loss (millions)	var aggregate loss (trillions)
Family Doctor	0.225	0.22496625	0.2925	0.5624429625
Surgeon	0.444	0.44367144	1.9536	22.5133190784
Nurse	3.264	3.25956096	1.632	1.33713024
Total			3.8781	24.4128922809

Using a Gamma approximation, the method of moments gives the following parameters:

$$\begin{aligned}\alpha\theta &= 3.8781\\ \alpha\theta^2 &= 24.4128922809\\ \theta &= \frac{24.4128922809}{3.8781} = 6.29506518164\\ \alpha &= \frac{3.8781}{6.29506518164} = 0.616053986432\end{aligned}$$

The expected reinsurance payment in millions is therefore

$$\begin{aligned}
\int_a^\infty (x-a)\frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)}\,dx &= \int_a^\infty \frac{x^{\alpha}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)}\,dx - a\int_a^\infty \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)}\,dx \\
&= \alpha\theta\int_a^\infty \frac{x^{\alpha}e^{-\frac{x}{\theta}}}{\theta^{\alpha+1}\Gamma(\alpha+1)}\,dx - a\int_a^\infty \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)}\,dx \\
&= 0.561106072057
\end{aligned}$$

The expected square of the reinsurance payment in trillions is

$$\begin{aligned}
&\int_a^\infty (x-a)^2\frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)}\,dx \\
&= \int_a^\infty \frac{x^{\alpha+1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)}\,dx - 2a\int_a^\infty \frac{x^{\alpha}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)}\,dx + a^2\int_a^\infty \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)}\,dx \\
&= \alpha(\alpha+1)\theta^2\int_a^\infty \frac{x^{\alpha+1}e^{-\frac{x}{\theta}}}{\theta^{\alpha+2}\Gamma(\alpha+2)}\,dx - 2a\alpha\theta\int_a^\infty \frac{x^{\alpha}e^{-\frac{x}{\theta}}}{\theta^{\alpha+1}\Gamma(\alpha+1)}\,dx + a^2\int_a^\infty \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)}\,dx \\
&= 6.42318008585
\end{aligned}$$

Thus the standard deviation in millions, is $\sqrt{6.42318008585 - 0.561106072057^2} = 2.47150562649$. The premium is therefore $561106.072057 + 2471505.62649 = \$3,032,611.70$.

2. *An insurance company is modelling claim data as following a log-normal distribution with $\sigma = 1$. It collects the following sample of claims:*

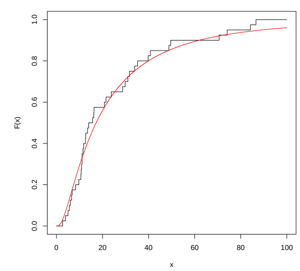
```
2.7 3.9 5.0 5.6 6.0 6.5 6.8 8.3 9.7 10.6 10.7 10.9
11.1 11.1 11.5 11.8 12.6 12.7 13.5 14.0 15.7 16.2 16.3
20.8 21.5 23.8 28.7 29.8 31.0 31.7 33.9 35.2 39.8 40.8
48.8 49.6 70.6 74.1 84.2 86.6
```

```
X<- c(2.7,3.9,5.0,5.6,6.0,6.5,6.8,8.3,9.7,10.6,10.7,
10.9,11.1,11.1,11.5,11.8,12.6,12.7,13.5,14.0,15.7,16.2,
16.3,20.8,21.5,23.8,28.7,29.8,31.0,31.7,33.9,35.2,39.8,
40.8,48.8,49.6,70.6,74.1,84.2,86.6)
```

The MLE for μ is 2.845. Graphically compare this empirical distribution with the best log-normal distribution with $\sigma = 1$. Include the following plots:

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

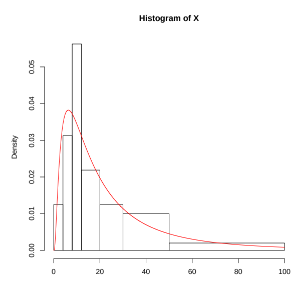
```
x<-(0:100000)/1000
mu<-2.845
FnX<-rowMeans(x%*%t(rep(1,40))>rep(1,100001)%*%t(X))
plot(x,FnX,type='l',xlab="x",ylab="F(x)")
FX<-pnorm(log(x)-mu)
FX[0]<-0 # The default formula works on my system, but in case it
         # fails on some systems, we explicitly set F_X(0).
points(x,FX,type='l',col="red")
```



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

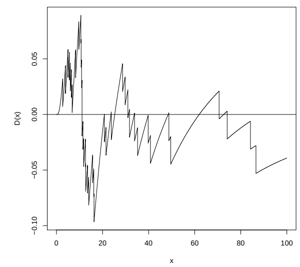
```
fx<-exp(-(log(x)-mu)^2/2)/(sqrt(2*pi)*x)
fx[1]<-0 # this density formula is undefined at x=0

hist(X,probability=TRUE,breaks=c(0,4,8,12,20,30,50,100))
  #These breaks produce a fairly smooth curve. Other choices are possible.
points(x,fx,type='l',col="red")
```



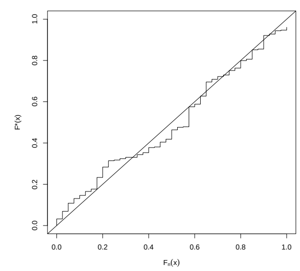
(c) A plot of $D(x)$ against x.

```
plot(x,FX-FnX,type='l',xlab="x",ylab="D(x)")
abline(h=0)
```



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

```
plot(FnX,FX,type='l',xlab=expression(F[n](x)),ylab="F*(x)",ylim=c(0,1))
abline(0,1)
```



3. *For the data in Question 2, calculate the following test statistics for the goodness of fit of the log-normal distribution with $\sigma = 1$ and $\mu = 2.845$:*

   (a) The Kolmogorov-Smirnov test.

X	$F_n(X^-)$	$F_n(X^+)$	$F^*(X)$	$-D(X^-)$	$D(X^+)$
2.7	0.000	0.025	0.03203099	0.0320309926	−0.007030993
3.9	0.025	0.050	0.06890135	0.0439013502	−0.018901350
5.0	0.050	0.075	0.10831069	0.0583106922	−0.033310692
5.6	0.075	0.100	0.13088161	0.0558816070	−0.030881607
6.0	0.100	0.125	0.14611538	0.0461153834	−0.021115383
6.5	0.125	0.150	0.16522750	0.0402274963	−0.015227496
6.8	0.150	0.175	0.17668371	0.0266837141	−0.001683714
8.3	0.175	0.200	0.23307899	0.0580789878	−0.033078988
9.7	0.200	0.225	0.28336497	0.0833649662	−0.058364966
10.6	0.225	0.250	0.31414113	0.0891411290	−0.064141129
10.7	0.250	0.275	0.31748036	0.0674803568	−0.042480357
10.9	0.275	0.300	0.32410970	0.0491097050	−0.024109705
11.1	0.300	0.325	0.33067325	0.0306732463	−0.005673246
11.1	0.325	0.350	0.33067325	0.0056732463	0.019326754
11.5	0.350	0.375	0.34360177	−0.0063982299	0.031398230
11.8	0.375	0.400	0.35312379	−0.0218762134	0.046876213
12.6	0.400	0.425	0.37778507	−0.0222149273	0.047214927
12.7	0.425	0.450	0.38079328	−0.0442067177	0.069206718
13.5	0.450	0.475	0.40426986	−0.0457301390	0.070730139
14.0	0.475	0.500	0.41841785	−0.0565821532	0.081582153
15.7	0.500	0.525	0.46361150	−0.0363884993	0.061388499
16.2	0.525	0.550	0.47608229	−0.0489177057	0.073917706
16.3	0.550	0.575	0.47853335	−0.0714666451	0.096466645
20.8	0.575	0.600	0.57532701	0.0003270146	0.024672985
21.5	0.600	0.625	0.58825285	−0.0117471539	0.036747154
23.8	0.625	0.650	0.62729048	0.0022904756	0.022709524
28.7	0.650	0.675	0.69563849	0.0456384940	−0.020638494
29.8	0.675	0.700	0.70867170	0.0336716970	−0.008671697
31.0	0.700	0.725	0.72206507	0.0220650714	0.002934929
31.7	0.725	0.750	0.72950503	0.0045050250	0.020494975
33.9	0.750	0.775	0.75124570	0.0012457040	0.023754296
35.2	0.775	0.800	0.76301855	−0.0119814473	0.036981447
39.8	0.800	0.825	0.79922800	−0.0007719985	0.025771999
40.8	0.825	0.850	0.80611872	−0.0188812798	0.043881280
48.8	0.850	0.875	0.85146339	0.0014633928	0.023536607
49.6	0.875	0.900	0.85519802	−0.0198019771	0.044801977
70.6	0.900	0.925	0.92102946	0.0210294573	0.003970543
74.1	0.925	0.950	0.92791205	0.0029120467	0.022087953
84.2	0.950	0.975	0.94387887	−0.0061211346	0.031121135
86.6	0.975	1.000	0.94698525	−0.0280147533	0.053014753

so the Kolmogorov-Smirnov statistic is 0.096466645.

(b) The Anderson-Darling test.

The Anderson-Darling statisic is given by

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

We calculate this for our data set

```
Asq<-40*(sum(((40:1)/40)^2*(c(0,log(pnorm(log(X[1:39])-mu,lower.tail=FALSE)))
-log(pnorm(log(X)-mu,lower.tail=FALSE))))+
sum(((1:40)/40)^2*(c(log(pnorm(log(X[2:40])-mu)),0)-log(pnorm(log(X)-mu))))-1)
```

This gives the Anderson-Darling statistic as 0.4102516.

(c) The chi-square test, dividing into the intervals 0–10, 10–20, 20–40 and more than 40.

We have the following table:

Interval	O	E	$\frac{(O-E)^2}{E}$
$[0,10)$	9	$40 \times 0.2937664 = 11.750656$	$\frac{(9-11.750656)^2}{11.750656} = 0.643888173591$
$[10,20)$	14	$40\,(0.5599065 - 0.2937664) = 10.645604$	$\frac{(14-10.645604)^2}{10.645604} = 1.0569595229$
$[20,40)$	10	$40\,(0.8006316 - 0.5599065) = 9.629004$	$\frac{(10-9.629004)^2}{9.629004} = 0.0142941089251$
$[40,\infty)$	7	$40(1-0.8006316) = 7.974736$	$\frac{(7-7.974736)^2}{7.974736} = 0.11914002792$
Total			1.83428183334

The Chi-squared statistic is 1.83428183334.

4. *For the data in Question 2, perform a likelihood ratio test to determine whether a log-normal distribution with fixed $\sigma = 1$, or a log-normal distribution with σ freely estimated is a better fit for the data. [For the general log-normal distribution, the MLE is $\sigma^2 = 0.7305957$ and $\mu = 2.845$.]*

The log-likelihood is given by

$$\sum_{i=1}^{40} -\log(x) - \log(\sigma) - \frac{1}{2}\log(2\pi) - \frac{(\log(X)-\mu)^2}{2\sigma^2}$$

We calculate this for the two parameter values

```
mu<-2.845
llGeneral<--20*log(2*pi)-20*log(0.7305957)-sum(log(X))-sum((log(X)-mu)^2)/(2*0.7305957)
llsigma1<--20*log(2*pi)-sum(log(X))-sum((log(X)-mu)^2/2)
```

Gives the log-likelihoods -164.266090367 and -165.155905697 respectively. Thus the log-likelihood ratio is $2(-164.266090367-(-165.155905697)) = 1.77963066$. This is compared to a chi-squared distribution with one degree of freedom, so the critical value, at the 5% significance level, is 3.841459, so we do not reject $\sigma = 1$.

5. *For the data in Question 2, use AIC and BIC to choose between a log-normal distribution with $\sigma = 1$ for the data and a transformed gamma distribution. [The MLE for the transformed gamma distribution is $\alpha = 0.02372$, $\theta = 87.77037$ and $\tau = 26.26$.]*

   The log-likelihood for the transformed gamma distribution is

$$40(\log(\tau) - \alpha\tau\log(\theta) - \log(\Gamma(\alpha))) + (\alpha\tau - 1)\sum(\log(X)) - \sum\left(\frac{X}{\theta}\right)^{\tau}$$

   We substitute the MLE for α, θ and τ to get that the log-likelihood is

$$40(\log(26.26) - 0.02372 \times 26.26\log(87.77037) - \log(\Gamma(0.02372))) + (0.02372 \times 26.26 - 1)\sum(\log(X)) - \sum\left(\frac{X}{87.77037}\right)^{26.26} = -173.8605$$

   The AIC for the log-normal with $\sigma = 1$ is $-165.155905697-1 = -166.155905697$, and the BIC is $-165.155905697 - \frac{1}{2}\log(40) = -167.000345424$

   For the transformed gamma distribution, the AIC is $-173.8605 - 3 = -176.8605$ and the BIC is $-173.8605 - \frac{3}{2}\log(40) = -179.393819181$. Thus the log-normal distribution is prefered by both AIC and BIC.

Standard Questions

6. *A health insurer divides insureds into three categories: non-smoker; occasional smoker; and heavy smoker. The number of claims made by an individual follows a negative binomial distribution with parameters r and β. It has the following portfolio of policies.*

Category	*Number insured*	*r*	*β of claim*	*mean claim*	*standard deviation*
non-smoker	*3,422*	*0.3*	*2.2*	*$860*	*$ 83,620*
occasional smoker	*1,053*	*0.9*	*2.4*	*$1,220*	*$113,190*
heavy smoker	*410*	*1.2*	*4.8*	*$1,740*	*$179,420*

   The insurance company models the aggregate losses as following a Pareto distribution with the correct mean and variance. It wants to buy stop-loss reinsurance for its policies. The reinsurance company uses the same

Pareto distribution to model aggregate losses and sets its premium at 125% of expected payments on the policy. The insurer sets the premium for the part of the losses that it covers as one standard deviation above expected aggregate payments it makes on the portfolio (and directly adds the reinsurer's premium to this). What attachment point for the reinsurance results in the smallest total premium for the policy? [You may need to numerically solve the derivative equal to zero. You may find the substitution $t = \frac{\theta}{\theta+a}$ helpful.]

We calculate the expectation and variance of aggregate claims

Category	$\mathbb{E}(N)$	$\mathrm{Var}(N)$	$\mathbb{E}(S)$	$\mathrm{Var}(S)$
Non-smoker	2258.52	7227.264	$2258.52 \times 860 = 1942327.2$	$7227.264 \times 860^2 + 2258.52 \times 83620^2 = 15,797,604,618,000$
Occasional Smoker	2274.48	7733.232	$2274.48 \times 1220 = 2774865.6$	$7733.232 \times 1220^2 + 2274.48 \times 113190^2 = 29,152,093,542,400$
Heavy Smoker	2361.6	13697.28	$2361.6 \times 1740 = 4109184$	$13697.28 \times 1740^2 + 2361.6 \times 179420^2 = 76,065,002,247,100$
Total			8826376.8	121014700407500

Setting these as the mean and variance of a Pareto distribution gives

$$\frac{\theta}{\alpha - 1} = 8826376.8$$
$$\frac{\alpha\theta}{(\alpha-1)^2(\alpha-2)} = 121014700407500$$
$$\frac{\alpha - 2}{\alpha} = \frac{8826376.8^2}{121014700407500} = 0.643764163799$$
$$\alpha = \frac{2}{1 - 0.643764163799} = 5.61425829958$$
$$\theta = 4.61425829958 \times 8826376.8 = 40727182.4046$$

For reinsurance with attachment point a, the expected payment is

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

The expected squared payment on the reinsurance policy is

$$
\begin{aligned}
\int_a^\infty 2(x-a)\left(\frac{\theta}{\theta+x}\right)^\alpha dx &= \int_{a+\theta}^\infty 2(u-a-\theta)\theta^\alpha u^{-\alpha}\,dx \\
&= 2\int_{a+\theta}^\infty \theta^\alpha(u^{1-\alpha}-(a+\theta)u^{-\alpha})\,dx \\
&= 2\theta^\alpha\left[\frac{u^{2-\alpha}}{(2-\alpha)}-(a+\theta)\frac{u^{1-\alpha}}{(1-\alpha)}\right]_{a+\theta}^\infty \\
&= 2\theta^\alpha(a+\theta)^{2-\alpha}\left(\frac{1}{(\alpha-2)}-\frac{1}{(\alpha-1)}\right) \\
&= 2\frac{\theta^\alpha}{(\alpha-1)(\alpha-2)(a+\theta)^{\alpha-2}}
\end{aligned}
$$

Let I be the portion of loss covered by the insurer, and R the portion covered by the reinsurer. The expected squared aggregate loss is

$$
\begin{aligned}
\mathbb{E}((I+R)^2) &= 2\frac{\theta^2}{(\alpha-1)(\alpha-2)} \\
\mathbb{E}(I)^2+\mathbb{E}(R^2)+2\mathbb{E}(RI) &= 2\frac{\theta^2}{(\alpha-1)(\alpha-2)}
\end{aligned}
$$

Now since whenever $R>0$, we have $I=a$, this gives

$$
\begin{aligned}
\mathbb{E}(R^2)+\mathbb{E}(I)^2+2a\mathbb{E}(R) &= 2\frac{\theta^2}{(\alpha-1)(\alpha-2)} \\
\mathbb{E}(I^2)+2\frac{\theta^\alpha}{(\alpha-1)(\alpha-2)(a+\theta)^{\alpha-2}}+\frac{2a\theta^\alpha}{(\alpha-1)(a+\theta)^{\alpha-1}} &= 2\frac{\theta^2}{(\alpha-1)(\alpha-2)} \\
\mathbb{E}(I^2) &= 2\frac{\theta^2}{(\alpha-1)(\alpha-2)}-2\frac{\theta^\alpha}{(\alpha-1)(\alpha-2)(a+\theta)^{\alpha-2}}-\frac{2a\theta^\alpha}{(\alpha-1)(a+\theta)^{\alpha-1}} \\
&= \frac{2\theta^2}{(\alpha-1)(\alpha-2)}\left(1-\left(\frac{\theta}{a+\theta}\right)^{\alpha-2}-\frac{(\alpha-2)a}{a+\theta}\left(\frac{\theta}{a+\theta}\right)^{\alpha-2}\right) \\
&= \frac{2\theta^2}{(\alpha-1)(\alpha-2)}\left(1-\left(\frac{\theta}{a+\theta}\right)^{\alpha-2}\left((\alpha-2)\frac{\theta}{a+\theta}-(\alpha-3)\right)\right)
\end{aligned}
$$

Substituting $t = \frac{\theta}{\alpha+\theta}$ we get

$$\begin{aligned}
\mathbb{E}(R) &= \frac{\theta}{\alpha-1}t^{\alpha-1} \\
\mathbb{E}(I) &= \frac{\theta}{\alpha-1} - \mathbb{E}(R) \\
&= \frac{\theta}{\alpha-1}(1-t^{\alpha-1}) \\
\mathbb{E}(I^2) &= \frac{2\theta^2}{(\alpha-1)(\alpha-2)}\left(1-t^{\alpha-2}\left((\alpha-2)t-(\alpha-3)\right)\right) \\
&= \frac{2\theta^2}{(\alpha-1)(\alpha-2)}\left(1-(\alpha-2)t^{\alpha-1}+(\alpha-3)t^{\alpha-2}\right) \\
\mathrm{Var}(I) &= \frac{2\theta^2}{(\alpha-1)(\alpha-2)}\left(1-(\alpha-2)t^{\alpha-1}+(\alpha-3)t^{\alpha-2}\right) - \frac{\theta^2}{(\alpha-1)^2}(1-t^{\alpha-1})^2 \\
&= \frac{\theta^2}{\alpha-1}\left(\frac{2}{\alpha-2} - 2t^{\alpha-1} + 2\frac{\alpha-3}{\alpha-2}t^{\alpha-2} - \frac{1}{\alpha-1} + \frac{2t^{\alpha-1}}{\alpha-1} - \frac{t^{2\alpha-2}}{\alpha-1}\right) \\
&= \frac{\theta^2}{\alpha-1}\left(\frac{\alpha}{(\alpha-1)(\alpha-2)} - 2\frac{\alpha-2}{\alpha-1}t^{\alpha-1} + 2\frac{\alpha-3}{\alpha-2}t^{\alpha-2} - \frac{t^{2\alpha-2}}{\alpha-1}\right)
\end{aligned}$$

The total premium is

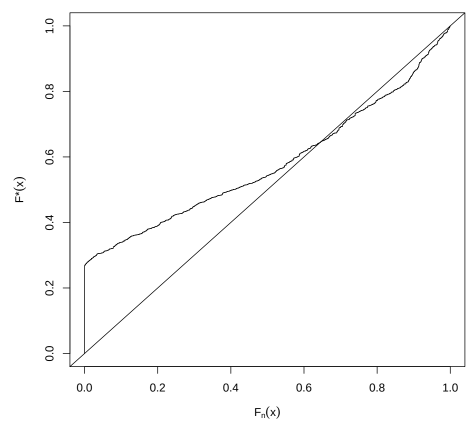
$$\mathbb{E}(I) + \sqrt{\mathrm{Var}(I)} + 1.25\mathbb{E}(R) = \frac{\theta}{\alpha-1} + 0.25\mathbb{E}(R) + \sqrt{\mathrm{Var}(I)}$$

We want to choose t to minimise this premium. This value is a solution to

$$\begin{aligned}
\frac{d}{dt}\left(\frac{\theta}{\alpha-1} + 0.25\mathbb{E}(R) + \sqrt{\mathrm{Var}(I)}\right) &= 0 \\
\frac{1}{2\sqrt{\mathrm{Var}(I)}}\frac{d\,\mathrm{Var}(I)}{dt} - 0.25\frac{dI}{dt} &= 0 \\
\frac{\theta^2}{2(\alpha-1)\sqrt{\mathrm{Var}(I)}}\left(2(\alpha-3)t^{\alpha-3} - 2(\alpha-2)t^{\alpha-2} - 2t^{2\alpha-3}\right) + \frac{\theta t^{\alpha-2}}{4} &= 0 \\
\theta\left(t^\alpha + (\alpha-2)t - (\alpha-3)\right) &= (\alpha-1)\frac{t}{4}\sqrt{\mathrm{Var}(I)} \qquad (1) \\
16\theta^2\left(t^\alpha + (\alpha-2)t - (\alpha-3)\right)^2 &= (\alpha-1)^2 t^2\,\mathrm{Var}(I) \\
&= t^2\theta^2\left(\frac{\alpha}{(\alpha-2)} - 2(\alpha-2)t^{\alpha-1} + 2\frac{(\alpha-1)(\alpha-3)}{\alpha-2}t^{\alpha-2} - t^{2\alpha-2}\right) \\
16\left(t^\alpha + (\alpha-2)t - (\alpha-3)\right)^2 &= \left(\frac{\alpha t^2}{(\alpha-2)} - 2(\alpha-2)t^{\alpha+1} + 2\frac{(\alpha-1)(\alpha-3)}{\alpha-2}t^{\alpha} - t^{2\alpha}\right) \qquad (2)
\end{aligned}$$

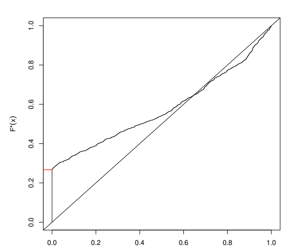
Numerically, we see that Equation (2) has two solutions: $t = 0.635009$ and $t = 0.739487$. For the first of these, we have that $t^\alpha + (\alpha-2)t - (\alpha-3) < 0$, so this is not a solution to Equation (1). Therefore, the solution is $t = 0.739487$. This corresponds to $a = \left(\frac{1}{0.739487} - 1\right)\theta = 0.352288816644\theta = \$14,197,263.7353$.

7. *An insurance company collects a sample of 700 past claims, and attempts to fit a distribution to the claims. Based on experience with other claims, the company believes that an inverse gamma distribution with $\alpha = 3.4$ and $\theta = 1,000$ may be appropriate to model these claims. It constructs the following p-p plot to compare the sample to this distribution:*



(a) What was the smallest data point of the sample?

From the graph, we see that the largest value for which $F_n(x) = 0$ is approximately $F^*(x) = 0.26$.



We solve $F^*(x) = 0.26$. Recall that the inverse of X is modelled as a gamma distribution with $\alpha = 3.4$. The 74th perecentile of a Gamma distribution with $\alpha = 3.4$ and $\theta = 1$ is 4.33. Thus the 26th percentile of the model distribution is $\frac{1000}{4.33} = 230.95$. Thus, this is approximately the smallest value observed in the sample.

[In fact, the smallest value in the sample used to generate this plot is 233.8657.]

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

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

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

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

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

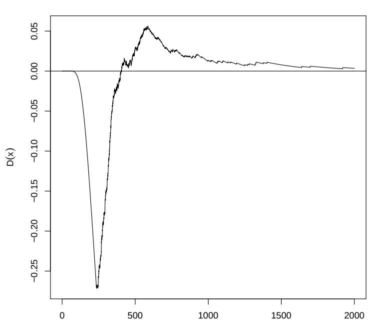
Justify your answer.

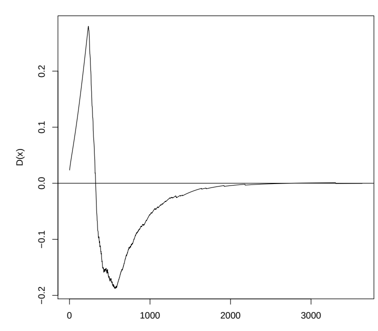
We have $F^*(x) > F_n(x)$ for small values of X, and $F^*(x) < F_n(x)$ for larger values of x. This means that the model assigns too much probability to small values, and too much probability to very large values, and therefore, too little probability to central values (iii).

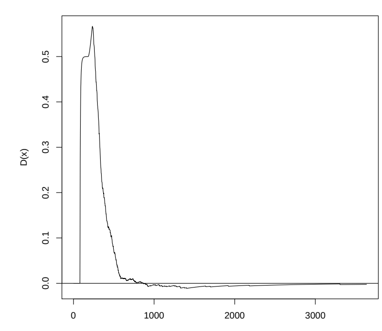
[The difference between $F^*(x)$ and $F_n(x)$ is much larger for small values than for large values, and we have $F^*(x) = F_n(x)$ for $x \approx 0.65$, so it is not unreasonable to decide that (ii) is a better description of the fit.]

(c) Which of the following plots is the $D(x)$ plot of this model on this data? Justify your answer.

(i) *(ii)* *(iii)*







Since $F^*(x) > F_n(x)$ for smaller values of x and $F^*(x) < F_n(x)$ for larger values, we have that $D(x) > 0$ for small x with $D(x) = 0.26$ when $x \approx 230$ and $D(x) < 0$ for large x. This is what we see in (i) but not in the other graphs. Therefore (i) must be the correct $D(x)$ plot.