ACSC/STAT 4703, Actuarial Models II

FALL 2024

Toby Kenney

Homework Sheet 4

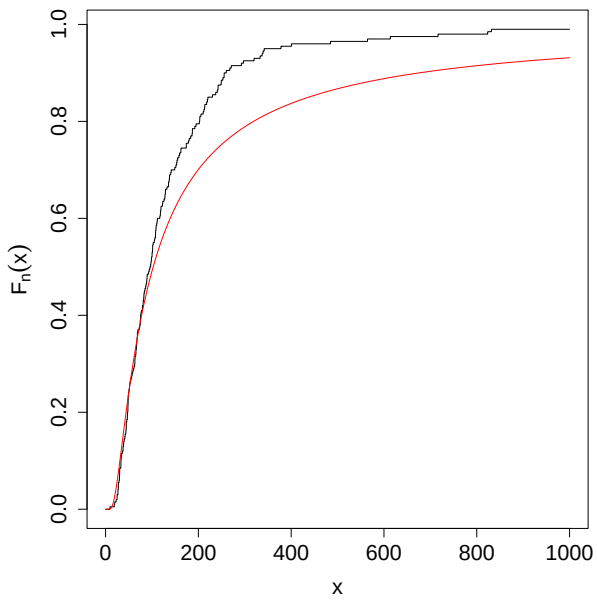
Model Solutions

Basic Questions

1. *The file* `HW4_data1.txt` *contains 200 i.i.d. samples of a random variable. An insurer is trying to model this random variable as following an inverse exponential distribution, as suggested by data sets from earlier years. Graphically compare this empirical distribution with the best inverse exponential distribution. From the data, they find that the MLE for* θ *is* $\theta = 71.1524$*. Include the following plots:*

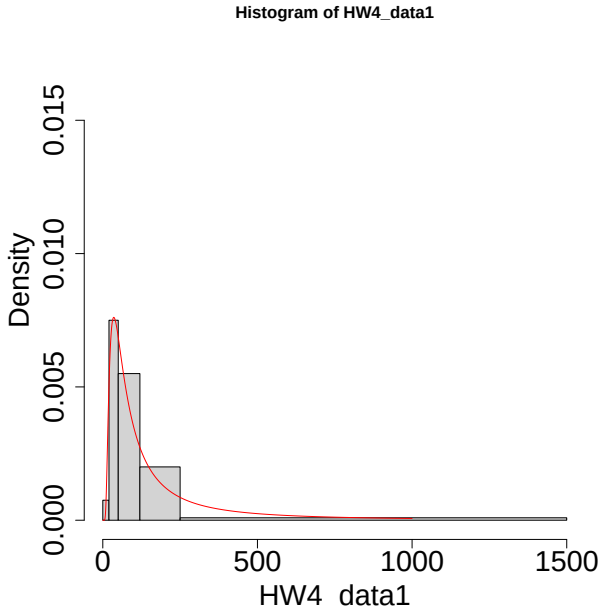
   (a) Comparisons of $F(x)$ *and* $F^*(x)$

```
### Fnx - count proportion of observations less than x.
x<-seq_len(10000)*0.1
theta<- 71.1524
alpha<-1
Fx<-rowMeans(x%*%t(rep(1,200))>rep(1,10000)%*%t(HW4_data1))
### Actually, can use Fx<-rowMeans(x>rep(1,10000)%*%t(HW4_data))
### Because R repeats vectors when comparing matrices of different sizes.

### Adjust margins to allow larger axis labels.
par(mar=c(4,5,1,1))
### Plot empirical cdf
plot(x,Fx,type='l',ylab=expression(F[n](x)),cex.axis=1.5,cex.lab=1.5)
### Plot model cdf
points(x,pgamma(theta/x,shape=alpha,lower.tail=FALSE),col="red",type='l')
```



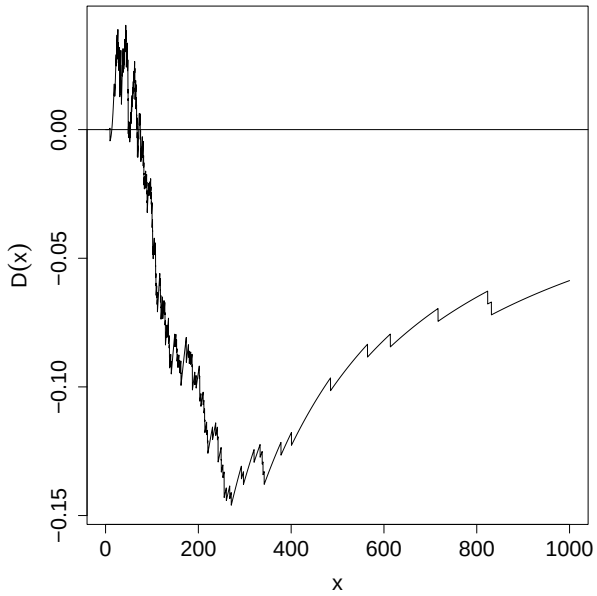
(b) Comparisons of $f(x)$ and $f^*(x)$

```
### Use built-in hist function
### Since I set unequal breaks, probability=TRUE is unnecessary.
par(mar=c(4,5,1,1))

hist(HW4_data1, probability=TRUE, breaks=c(0,20,50,120,250,1500),
     cex.axis=2,cex.lab=2,ylim=c(0,0.018))
### The default evenly spaced breaks put most of the data in the first
### bar, and thus do not give a great picture of the distribution.

### plot the model density on the same graph.
points(x,theta^alpha/x^(alpha+1)/gamma(alpha)*exp(-theta/x),type='l',col="red")
```



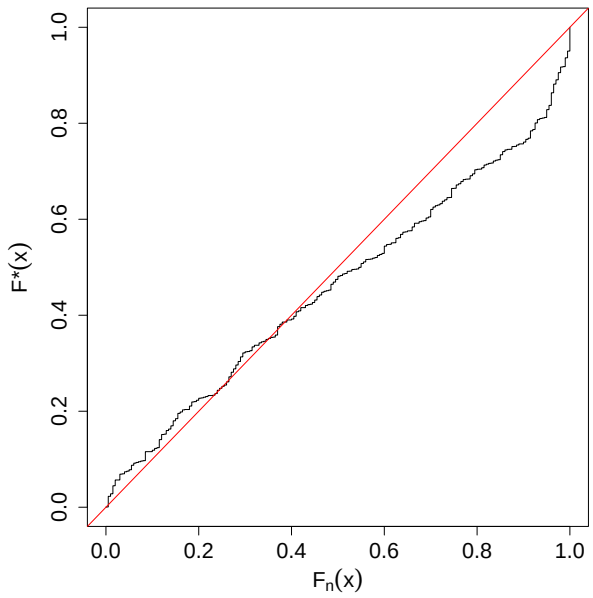
(c) A plot of $D(x)$ against x.

```
### Adjust margins to allow larger axis labels.
par(mar=c(4,5,1,1))
###
plot(x,pgamma(theta/x,shape=alpha,lower.tail=FALSE)-Fx,type='l',
     ylab=expression(D(x)),cex.axis=1.5,cex.lab=1.5)
### Plot the reference line
abline(h=0)
```



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

```
Fstar<-pgamma(theta/sort(HW4_data1),shape=alpha,lower.tail=FALSE)
Fstar_repeat<-c(0,rep(Fstar,each=2),1)
n<-length(HW4_data1)
Fn_lower_upper<-rep(c(0,seq_len(n)/n),each=2)

### Adjust margins to allow larger axis labels.
par(mar=c(4,5,1,1))
### Plot empirical cdf
plot(Fn_lower_upper,Fstar_repeat,type='l',ylab=expression(paste(F,"*")(x)),
     xlab=expression(F[n](x)),cex.axis=1.5,cex.lab=1.5)
### Plot model cdf
abline(0,1,col="red")
```



2. *For the data in* `HW4_data1.txt`*, calculate the following test statistics for the goodness of fit of the Inverse exponential distribution with θ estimated by MLE:*

   (a) The Kolmogorov-Smirnov test.

   Using the following code:

```
HW4_data1<-read.table("HW4_data1.txt")[[1]]
HW4_sorted<-sort(HW4_data1)
n<-length(HW4_sorted)
theta<- 71.1524

Fstari<-pgamma(theta/HW4_data1,shape=alpha,lower.tail=FALSE) # Model CDF
Fn.plus<-seq_len(n)/n # empirical CDF above
Fn.minus<-(seq_len(n)-1)/n # empirical CDF below

KS<-max(c(Fn.plus-Fstari,Fstari-Fn.minus))
```

   the Kolmogorov-Smirnov statistic is 0.146065, attained at the sample $x = 270.8$.

   (b) The Anderson-Darling test.

   We use the following code:

```
200*(sum(((200:0)/200)^2*(c(0,log(1-Fstari))-c(log(1-Fstari[seq_len(200)]),0)))+
    sum(((1:200)/200)^2*(c(log(Fstari[seq_len(199)+1]),0)-log(Fstari)))-1)
```

This gives the Anderson-Darling statistic as 5.904214.

(c) The chi-square test, dividing into the intervals 0–50,50–100,100–300 and more than 300.

The expected number of observations in the interval $[a, b]$ are 200 times The probability of the interval $[a, b]$. We use the following `R` code to make a table.

```
cut.CDF<-c(0,pgamma(theta/c(50,100,300),shape=1,lower.tail=FALSE),1)
Obs.freq<-table(cut(HW4_data1,breaks=c(0,50,100,300,6000),right=FALSE))
                                        # Observed frequencies
Exp.freq<-200*(cut.CDF[-1]-cut.CDF[-5]) #Expected Frequencies
cbind(Obs.freq,Exp.freq,(Obs.freq-Exp.freq)^2/Exp.freq)
sum((Obs.freq-Exp.freq)^2/Exp.freq)
```

This gives the following table:

Interval	E	O	$\frac{(O-E)^2}{E}$
$[0, 50)$	48	48.19568	0.001
$[50, 100)$	56	49.98342	0.724
$[100, 300)$	81	59.59160	7.691
$[300, \infty)$	15	42.22930	17.557
Total			25.973

The Chi-squared statistic is 25.973.

3. *For the data in* `HW4_data1.txt`*, perform a likelihood ratio test to determine whether an inverse exponential distribution, or an inverse transformed gamma distribution with α, τ and θ freely estimated is a better fit for the data. [For the inverse transformed gamma distribution, the MLE is $\alpha = 8.603$, $\tau = 0.4237$ and $\theta = 13615$.]*

The log-likelihood for the inverse transformed gamma distribution is given by

$$\sum_{i=1}^{200} \log(\tau) + \tau\alpha \log(\theta) - \left(\frac{\theta}{x_i}\right)^{\tau} - (\alpha\tau + 1)\log(x_i) - \log(\Gamma(\alpha))$$

We calculate this for the parameter values

```
alpha<-8.603
tau<-0.4237
theta<-13615
200*(log(tau)+tau*alpha*log(theta)-log(gamma(alpha)))-
    sum(theta^tau/HW4_data1^tau)-(tau*alpha+1)*sum(log(HW4_data1))
```

giving the log-likelihood -1160.1.

The log-likelihood for the inverse exponential distribution is $n\log(\theta) - 2\sum_{i=1}^{n}\log(x_i) - \sum_{i=1}^{n}\frac{\theta}{x_i}$ which gives the log-likelihood -1178.8. Thus the log-likelihood ratio is $2(-1160.1-(-1178.8)) = 37.4$. This is compared to a chi-squared distribution with two degrees of freedom, so the critical value, at the 5% significance level, is 5.991465, so we reject the inverse exponential distribution.

4. *For the data in* `HW4_data1.txt`*, use AIC and BIC to choose between an inverse exponential distribution and an inverse Pareto distribution. [The MLE for the inverse Pareto distribution is* $\tau = 0.925$ *and* $\theta = 129.9$*.]*

   The log-likelihood for the inverse Pareto distribution is

   $$\sum_{i=1}^{200}\log(\tau) + \log(\theta) + (\tau-1)\log(x_i) - (\tau+1)\log(\theta+x_i)$$

   We substitute the MLE for τ and θ to calculate the log-likelihood:

```
tau<-0.925
theta<-129.9
200*log(tau)+200*log(theta)+(tau-1)*sum(log(HW4_data1))-
    (tau+1)*sum(log(theta+HW4_data1))
```

   This gives the log-likelihood as -1230.478

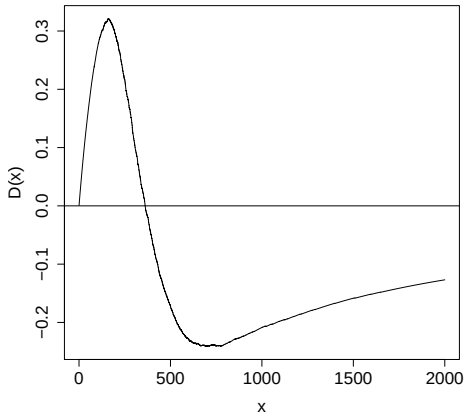
   The AIC for the inverse exponential distribution is $-1178.815 - 1 = -1179.815$, and the BIC is $-1178.815 - \frac{1}{2}\log(200) = -1181.46415868$

   For the inverse Pareto distribution, the AIC is $-1230.478-2 = -1232.478$ and the BIC is $-1230.478-\log(200) = -1235.77631737$. Thus the inverse exponential distribution is prefered by both AIC and BIC.

Standard Questions

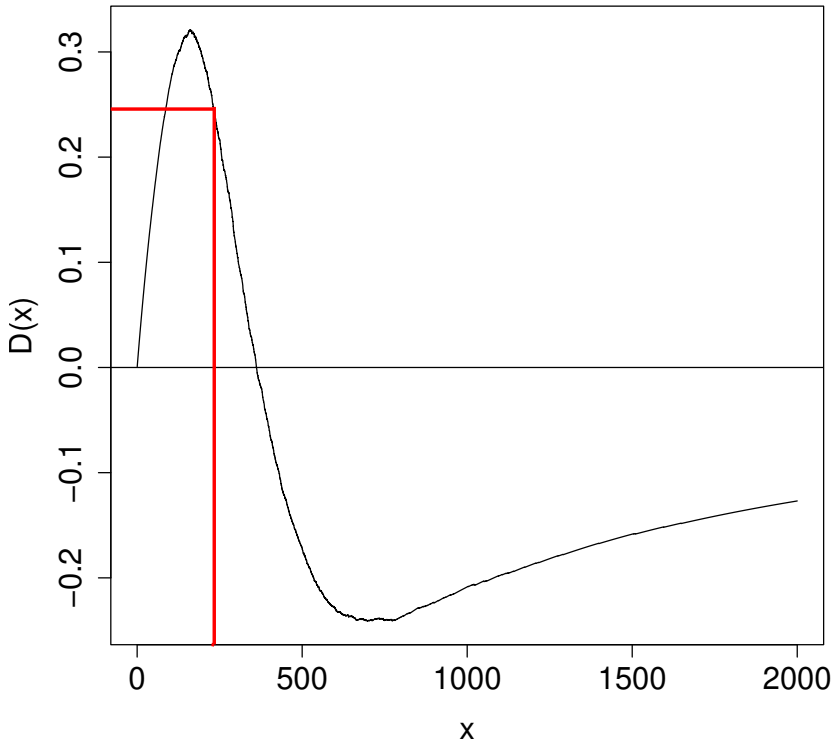
5. *An insurance company collects a sample of 2741 past claims, and attempts to fit a distribution to the claims. Based on experience with other claims, the actuary believes that a Pareto distribution may be appropriate*

to model these claims. She fits the MLE parameters $\alpha = 0.9085505$ *and* $\theta = 230.6825$ *and constructs the following plot* $D(x) = F^*(x) - F_n(x)$ *for this distribution and data.*



(a) How many data points in the sample were less than 240?

We have that $F^*(240) = 1 - \left(\frac{230.6825}{230.6825+240}\right)^{0.9085505} = 0.476869870677.$



From the graph, we read $D(240) \approx 0.24$, so $F_n(240) \approx 0.476869870677 - 0.24 = 0.236869870677$. So there are approximately $2741 \times 0.236869870677 = 649.260315526$ samples less than 240 in the dataset. [In fact, there are 674 samples less than 240 in the data set.]

(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

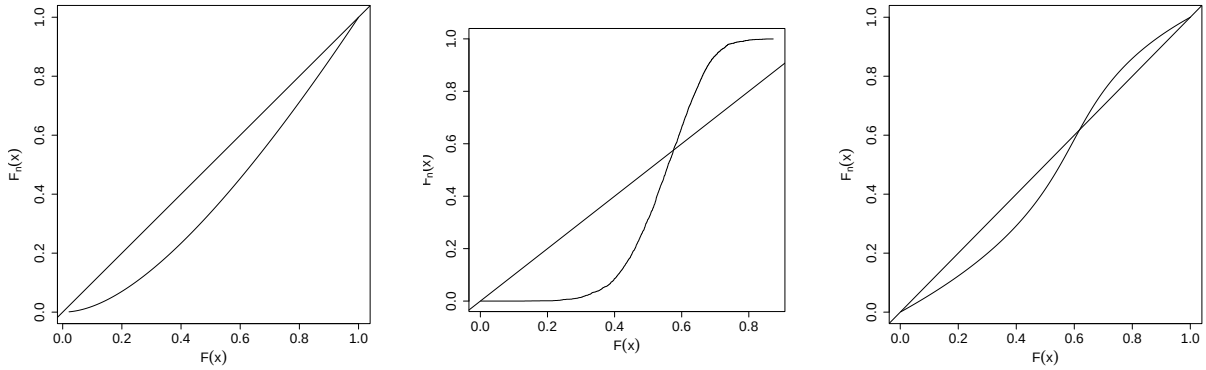
and too little probability to tail values.

Justify your answer.

We see that $D(x) > 0$ for $x < 300$ and $D(x) < 0$ for $x > 400$, so the $F^*(x) > F_n(x)$ for small x, and $F^*(x) < F_n(x)$ for large values of x. This means that the Pareto distribution assigns too much probability to tail values and too little probability to central values, so (iii) best describes the fit.

(c) Which of the following plots is a p-p plot for this model on this data? Justify your answer.

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



We know that $F^*(x) > F_n(x)$ for small x and $F^*(x) < F_n(x)$ for large x, so (i) is not correct. Between (ii) and (iii), we see that the largest value of $D(x)$ is slightly above 0.3. For the plot in (iii), the largest value is much smaller. Thus (ii) must be the correct plot.