ACSC/STAT 4703, Actuarial Models II

FALL 2023

Toby Kenney

Homework Sheet 4

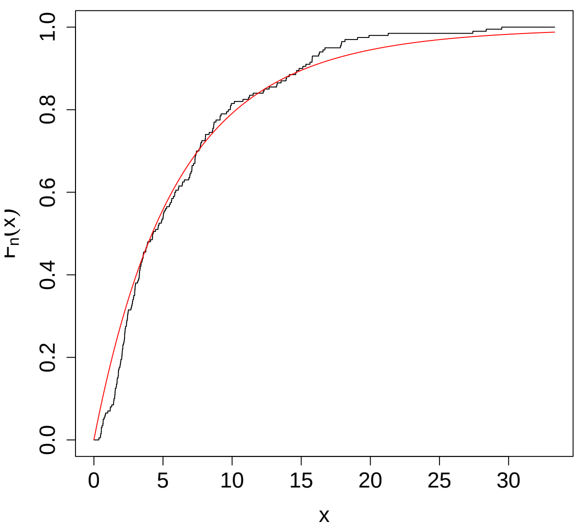
Model Solutions

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 a Pareto distribution with $\alpha = 9$, as suggested by data sets from earlier years. Graphically compare this empirical distribution with the best Pareto distribution with $\alpha = 9$. From the data, they find that the MLE for θ is $\theta = 52.61$. Include the following plots:*

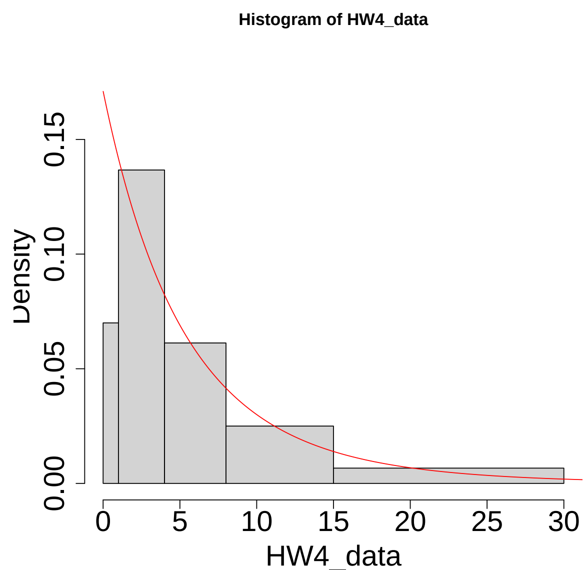
   (a) Comparisons of $F(x)$ and $F^*(x)$

```
### Fnx - count proportion of observations less than x.
x<-seq_len(10000)*0.0035
theta<-52.61
Fx<-rowMeans(x%*%t(rep(1,200))>rep(1,10000)%*%t(HW4_data))
### 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,1-(theta/(theta+x))^9,col="red",type='l')
```



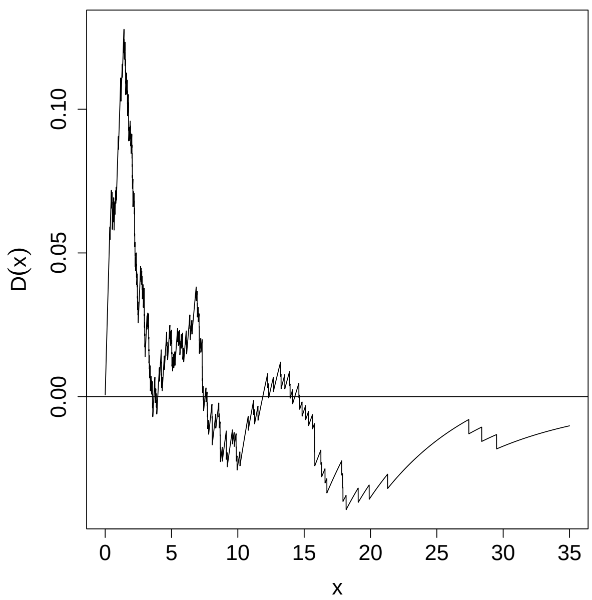
(b) Comparisons of $f(x)$ and $f^*(x)$

```
### Use built-in hist function
### Since I set unequal breaks, probability=TRUE is unnecessary.
hist(HW4_data,probability=TRUE,breaks=c(0,1,4,8,15,30),cex.axis=2,cex.lab=2,ylim=c(0,0.18
### The default evenly spaced breaks cover up the small first bar, and
### produce some bars based on a very small number of points.

### plot the model density on the same graph.
points(x,9*52.61^9/(52.61+x)^10,type='l',col="red")
```



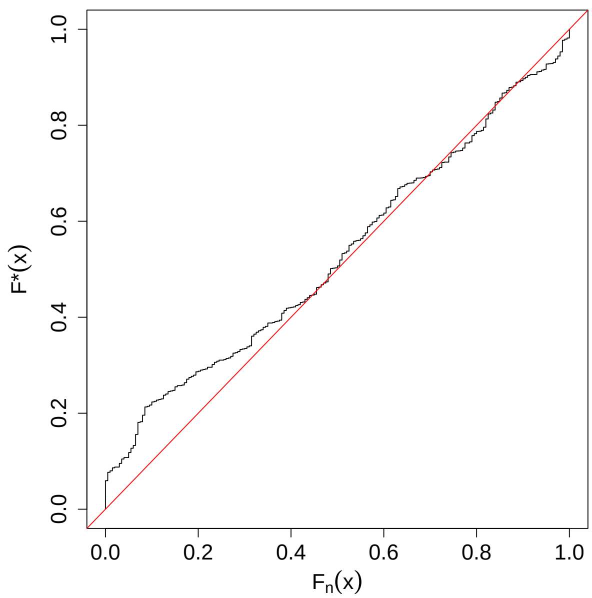
(c) A plot of $D(x)$ against x.

```
### Adjust margins to allow larger axis labels.
par(mar=c(4,5,1,1))
### Plot empirical cdf
plot(x,1-(theta/theta+x)^alpha-Fx,type='l',ylab=expression(D(x)),cex.axis=1.5,cex.lab=1.5
### Plot model cdf
abline(h=0)
```



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

```
Fstar<-1-(theta/(theta+sort(HW4_data)))^9
Fstar_repeat<-c(0,rep(Fstar,each=2),1)
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=expressio
### 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 Pareto distribution with* $\alpha = 9$ *and* θ *estimated by MLE:*

   (a) The Kolmogorov-Smirnov test.

   Using the following code:

```
HW4_data<-read.table("HW4_data1.txt")
HW4_sorted<-sort(HW4_data[[1]])
n<-length(HW4_sorted)
theta <-52.61

Fstari<-1-(theta/(theta+HW4_sorted))^9 # 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 0.1281354, attained at the sample $x = 1.42$.

   (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 3.671391.

(c) The chi-square test, dividing into the intervals 0–1,1–5,5–10 and more than 10.

The probability of the interval $[a, b]$ is $\left(\frac{\theta}{\theta+a}\right)^9 - \left(\frac{\theta}{\theta+b}\right)^9$. The expected number of observations are 200 times this. We use the following `R` code to make a table.

```
cut.Surv<-c((theta/(theta+c(0,1,5,10)))^9,0)
Obs.freq<-table(cut(HW4_data,breaks=c(0,1,5,10,1000),right=FALSE))# Observed frequencies
Exp.freq<-200*(cut.Surv[-5]-cut.Surv[-1]) #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	$\frac{(O-E)^2}{E}$
$[0, 1)$	31.17668	13	10.5973950
$[1, 5)$	80.48201	94	2.2705190
$[5, 10)$	46.57257	56	1.9083431
$[10, \infty)$	41.76874	37	0.5444475
Total			15.3207

The Chi-squared statistic is 15.3207.

3. *For the data in* `HW4_data1.txt`, *perform a likelihood ratio test to determine whether a Pareto distribution with fixed $\alpha = 9$, or a generalised Pareto distribution with α, τ and θ freely estimated is a better fit for the data. [For the generalised Pareto distribution, the MLE is $\alpha = 5.6701$, $\tau = 1.86747$ and $\theta = 15.89494$.]*

The log-likelihood is given by

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

We calculate this for the two parameter values

```
alpha<-5.6701
tau<-1.86747
theta<-15.89494
200*(log(gamma(alpha+tau))-log(gamma(alpha))-log(gamma(tau))+alpha*log(theta))+
    (tau-1)*sum(log(HW4_data))-(alpha+tau)*sum(log(theta+HW4_data))
```

Gives the log-likelihoods -559.8156 and -571.4705 respectively. Thus the log-likelihood ratio is $2(-559.8156 - (-571.4705)) = 23.3098$. 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 $\alpha = 9, \tau = 1$.

4. *For the data in* `HW4_data1.txt`*, use AIC and BIC to choose between a Pareto distribution with $\alpha = 9$ for the data and a transformed gamma distribution. [The MLE for the transformed gamma distribution is $\alpha = 0.883801$, $\tau = 1.304570$ and $\theta = 7.650101$.]*

   The log-likelihood for the transformed gamma distribution is

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

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

```
200*log(tau)+tau*alpha*sum(log(HW4_data))-200*tau*alpha*log(theta)-
    sum((HW4_data/theta)^tau)-sum(log(HW4_data))-200*log(gamma(alpha))
```

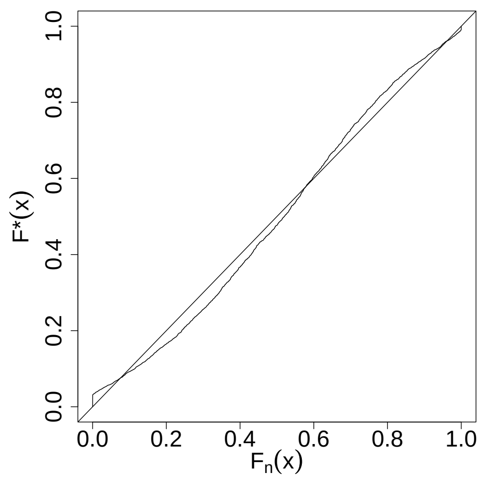
   This gives the log-likelihood as -564.0346

   The AIC for the Pareto distribution with $\alpha = 9$ is $-571.4705 - 1 = -572.4705$, and the BIC is $-571.4705 - \frac{1}{2}\log(200) = -574.119658683$

   For the transformed gamma distribution, the AIC is $-564.0346 - 3 = -567.0346$ and the BIC is $-564.0346 - \frac{3}{2}\log(200) = -571.98207605$. Thus the transformed gamma distribution is prefered by both AIC and BIC.

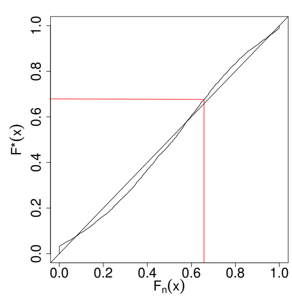
Standard Questions

5. *An insurance company collects a sample of 3,900 past claims, and attempts to fit a distribution to the claims. Based on experience with other claims, the actuary believes that a log-normal distribution may be appropriate to model these claims. She fits the MLE parameter $\mu = 0.4373128$ and $\sigma^2 = 0.3691496$ and constructs the following p-p plot of the distribution and data.*



(a) How many data points in the sample were more than 2?

We have that $F^*(2) = \Phi\left(\frac{\log(2)-0.4373128}{\sqrt{0.3691496}}\right) = 0.6631492$. From the graph, we read $F_n(2) \approx 0.64$.



So there are approximately $3900 \times 0.36 = 1404$ samples larger than 2 in the dataset. [In fact, there are 1386 samples larger than 2 in the data set.]

(b) Which of the following statements best describes the fit of the log-normal distribution to the data:

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

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

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

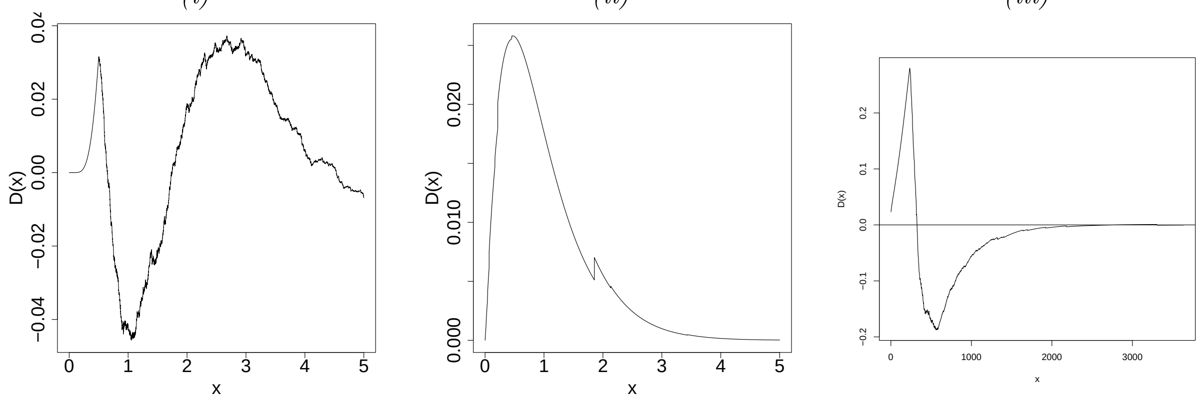
(iv) The log-normal distribution assigns too much probability to central values and too little probability to tail values.

Justify your answer.

We see that $F_n(x) > F^*(x)$ for $0.08 < F^*(x) < 0.57$ and $F_n(x) < F^*(x)$ for $0.57 < F^*(x) < 0.94$. Thus suggests that $F^*(x)$ grows much faster than $F_n(x)$ between these values, so (iv) F^* assigns too much probability to central values, and too little to tail values. [However, the very extreme tails tell a different story, so you could argue for (iii)]

(c) Which of the following plots shows $D(x) = F^*(x) - F_n(x)$ for this model on this data? Justify your answer.

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



Since $F^*(x) < F_n(x)$ for smaller (but not very small) values of x and $F^*(x) > F_n(x)$ for larger values, we expect $D(x)$ to be negative for small values of x and positive for larger values of x. Only (i) shows this pattern, so (i) must be the correct plot.