ACSC/STAT 4703, Actuarial Models II

FALL 2023

Toby Kenney

Practice Midterm Examination

Model Solutions

This Sample examination has more questions than the actual midterm, in order to cover a wider range of questions. Estimated times are provided after each question to help your preparation.

Here are some values of the Gamma distribution function with $\theta = 1$ that may be needed for this examination:

x	α	$F(x)$
245	255	0.2697208
$\left(\frac{7.5}{12}\right)^3$	$\frac{4}{3}$	0.1117140
$\left(\frac{9.5}{12}\right)^3$	$\frac{4}{3}$	0.2507382
2.5	1	0.917915
2.5	2	0.7127025
2.5	3	0.4561869
2.5	4	0.2424239
0.3542	3	0.005692012
5.6458	3	0.9202284

1. *An insurer is assessing a model. Under the model, a certain statistic X should follow a gamma distribution with parameters θ_1 and $\alpha = 3$, and another statistic Y should follow a Weibull distribution with $\tau = 2$ and scale parameter θ_2. They compute the statistic $\frac{X}{\theta_1} + \left(\frac{Y}{\theta_2}\right)^2$. What is the probability that this statistic exceeds 7? [10 mins]*

   $\frac{X}{\theta_1}$ should follow a gamma distribution with $\theta = 1$ and $\alpha = 3$. The density function is therefore $f(x) = \frac{x^2e^{-x}}{2}$. The probability that $\frac{X}{\theta_1} > 7$ is therefore 0.02963616. $\frac{Y}{\theta_2}$ should follow a Weibull distribution with $\theta = 1$ and $\tau = 2$. The probability that the square of this exceeds $7 - x$ is therefore $e^{-\left(\sqrt{7-x}\right)^2}$. Thus, the probability that $\frac{X}{\theta_1} + \left(\frac{Y}{\theta_2}\right)^2$ exceeds 7 is

   $$0.02963616 + \int_0^7 \frac{x^2e^{-x}}{2} e^{-(7-x)}\,dx = 0.02963616 + \frac{e^{-7}}{2}\int_0^7 x^2\,dx = 0.02963616 + \frac{e^{-7}}{2}\left[\frac{x^3}{3}\right]_0^7 = 0.02963616 + \frac{7^3}{6}e^{-7} = 0.0817654$$

2. *An insurer models claims as following a Pareto distribution with $\theta = 2000$ and α varying between individuals. They model $\alpha = 2 + 2A$ where A follows a gamma distribution with $\alpha = 3$ and $\theta = 1$. What is the VaR at the 0.95 level of the loss distribution for a random individual?*

   (i) 1243

   (ii) 8445

   (iii) 9290

   (iv) 15919

For an individual with $A = a$, the probability that the loss exceeds x is $S(x) = \left(\frac{2000}{2000+x}\right)^{2+2a}$. Let $p = \frac{2000}{2000+x}$. The overall probability of a loss exceeding x is $\mathbb{E}\left(p^{2+2A}\right) = p^2\mathbb{E}\left(p^{2A}\right)$. For the gamma distribution, we have

$$\begin{aligned}\mathbb{E}\left(p^{2A}\right) &= \frac{1}{2}\int_0^\infty a^2 e^{-a} e^{2\log(p)a}\, da \\ &= \frac{1}{2}\int_0^\infty a^2 e^{-(1-2\log(p))a}\, da \\ &= (1-2\log(p))^{-3}\end{aligned}$$

Thus, we need to solve $p^2(1-2\log(p))^{-3} = 0.05$.

We try the values given in the question:

	x	p	$p^2(1-2\log(p))^3$
(i)	1243	0.6167129	0.05000
(ii)	8445	0.1914792	0.00046
(iii)	9290	0.1771479	0.00035
(iv)	15919	0.1116134	0.00008

We see that (i) 1243 is the VaR.

3. *An insurance company models aggregate losses following a Pareto distribution with $\alpha = 8$ and $\theta = 9600$ for $x < \$50,000$ and a Pareto distribution with $\alpha = 3$ for $x > \$50,000$. The probability that a loss exceeds \$50,000 is 0.00001. The scale parameters of the Pareto distributions are chosen so that the overall density function is continuous. What is the expected aggregate loss? [15 mins]*

Let the scale parameter of the Pareto distribution fo $x > \$100,000$ be θ. Under the first Pareto distribution, the probability that a loss is less than \$50,000 is $1 - \left(\frac{9600}{59600}\right)^8 = 0.999999546895$, so this Pareto density is scaled by a factor $\frac{0.99999}{0.999999546895} = 0.999990453101$. Similarly, the second Pareto density is scaled by a factor $\frac{0.00001(50000+\theta)^3}{\theta^3}$. The scaled density of the first Pareto distribution at $x = 50000$ is given by $0.999990453101\frac{8\times 9600^8}{(59600)^9} = 6.08190129348\times 10^{-11}$ and the scaled density of the second Pareto distribution is given by $\frac{0.00001(100000+\theta)^3}{\theta^3}\frac{3\theta^3}{(\theta+100000)^4} = \frac{0.00003}{\theta+50000}$. For the spliced distribution to be continuous, these must be equal. That is

$$\begin{aligned}\frac{0.00003}{\theta+50000} &= 6.08190129348\times 10^{-11} \\ \theta + 50000 &= \frac{0.00003}{6.08190129348\times 10^{-11}} \\ &= 493266.801817 \\ \theta &= 443266.801817\end{aligned}$$

The conditional expectation of $X|X > 50000$ is

$$\left(\frac{50000+\theta}{\theta}\right)^3\int_{50000}^\infty \left(\frac{\theta}{\theta+x}\right)^3 dx = (50000+\theta)^3\int_{50000+\theta}^\infty u^{-3}\, du = \frac{(50000+\theta)}{2} = \frac{493266.801817}{2} = 246633.400909$$

The conditional expectation of $X|X < 50000$ is

$$\begin{aligned}
\frac{1}{0.999999546895}\int_0^{50000}\frac{8\times 9600^8 x}{(9600+x)^9}\,dx &= 8.00000362484\times 9600^8\int_0^{50000}(9600+x)^{-8}-9600(9600+x)^{-9}\,dx\\
&= 8.00000362484\times 9600^8\int_{9600}^{59600}u^{-8}-9600u^{-9}\,dx\\
&= 8.00000362484\times 9600^8\left[9600\frac{u^{-8}}{8}-\frac{u^{-7}}{7}\right]_{9600}^{59600}\\
&= 8.00000362484\times 9600\left(\frac{1}{56}+\frac{9600^8}{8\times 59600^8}-\frac{9600^7}{7\times 59600^7}\right)\\
&= 1371.40267965
\end{aligned}$$

The overall expectation is therefore

$$0.99999\times 1371.40267965+0.00001\times 246633.400909 = \$1,373.86$$

4. *An insurance company has the following data on its policies:*

Policy limit	*Losses Limited to*			
	20,000	*50,000*	*100,000*	*500,000*
20,000	*1,400,000*			
50,000	*7,540,000*	*8,010,000*		
100,000	*22,600,000*	*24,100,000*	*28,700,000*	
500,000	*5,900,000*	*6,220,000*	*6,650,000*	*6,920,000*

Use this data to calculate the ILF from \$20,000 to \$500,000 using

(a) The direct ILF estimate. [5 mins]

The direct ILF estimate is $\frac{6920000}{5900000} = 1.17288135593$.

(b) The incremental method. [5 mins]

Using the incremental method the ILFs are:

\$20,000–\$50,000	$\frac{8010000+24100000+6220000}{7540000+22600000+5900000} = 1.06354051054$
\$50,000–\$100,000	$\frac{28700000+6650000}{24100000+6220000} = 1.16589709763$
\$100,000–\$500,000	$\frac{6920000}{6650000} = 1.04060150376$

So the ILF is $1.06354051054\times 1.16589709763\times 1.04060150376 = 1.29032379813$.

5. *An insurance company charges a risk charge equal to the square of the average loss amount, divided by 100,000. It has the following data on a set of 1,200 claims from policies with limit \$1,000,000.*

Losses Limited to	*50,000*	*100,000*	*500,000*	*1,000,000*
Total claimed	*16,700,000*	*20,880,000*	*27,030,000*	*32,410,000*

Calculate the ILF from \$100,000 to \$1,000,000. [10 mins]

For limit \$100,000, the expected loss amount is $\frac{20880000}{1200} = 17400$, and the risk charge is $\frac{17400^2}{100000} = 3027.6$. The premium is therefore $17400+3027.6 = 20427.6$. For limit \$1,000,000, the expected loss amount is $\frac{32410000}{1200} = 27008.3333333$,

and the risk charge is $\frac{27008.3333333^2}{100000} = 7294.50069443$, so the premium is 27008.3333333+7294.50069443 = 34302.8340277. The ILF is therefore $\frac{34302.8340277}{20427.6} = 1.6792395596$.

6. *An insurer calculates the ILF on the pure premium from \$1,000,000 to \$2,000,000 on a particular policy is 1.092. A reinsurer offers excess-of-loss reinsurance of \$1,000,000 over \$1,000,000 for a loading of 30%. The original insurer uses a loading of 20% on policies with limit \$1,000,000. If the insurer buys the excess-of-loss reinsurance, what is the loading on its premium for policies with a limit of \$2,000,000? [10 mins]*

   Let m be the expected loss on the policy with limit \$1,000,000. With a 20% loading, the insurer charges $1.2m$ for the insurance. The expected payment on the reinsurance is $1.092m - m = 0.092m$. With a loading of 30%, the cost of the reinsurance is $0.092m \times 1.3 = 0.1196m$, so the total cost with a limit of \$2,000,000 is $1.2m + 0.1196m = 1.3196m$, and the expected payment is $1.092m$, so the loading is $\frac{1.3196m}{1.092m} - 1 = 20.842490842\%$.

7. *An insurer models a loss as following a Weibull distribution with $\tau = 4$ and $\theta = 100$. What are the parameters c_n and d_n that make the distribution of $\frac{M_n - d_n}{c_n}$ converge, where M_n are block maxima of a block of n samples, and what is the limiting distribution? [15 mins]*

   As the Weibull distribution is unbounded and has all finite moments, we know that the limiting distribution must be a Gumbel distribution. The limiting distribution has distribution function given by

$$-\log(H(x)) = \lim_{n\to\infty} nS(c_n x + d_n)$$

   The survival function of the Weibull distribution is $S(x) = e^{-\left(\frac{x}{100}\right)^4}$, so we want to find c_n and d_n such that

$$\lim_{n\to\infty} ne^{-\left(\frac{c_n x + d_n}{100}\right)^4} = e^{-x}$$

   For $x = 0$, this becomes

$$\lim_{n\to\infty} ne^{-\left(\frac{d_n}{100}\right)^4} = 1$$

   which will hold if $d_n = 100\log(n)^{\frac{1}{4}}$. We now need to select c_n so that the limit converges for every x. We see that if c_n is much larger than d_n, then the limit will converge to 0 for $x > 0$. Instead, we will need $\frac{c_n}{d_n} \to 0$. In this case, we will have $(c_n x + d_n)^4 \approx d_n^4 + 4d_n^3 c_n x$. Using this approximation, we need to ensure convergence of

$$\lim_{n\to\infty} ne^{-\frac{d_n^4}{100^4}} e^{-\frac{4d_n^3 c_n x}{100^4}} = e^{-\frac{4d_n^3 c_n x}{100^4}}$$

   This will happen if $4d_n^3 c_n$ converges to a constant nonzero limit. Thus, we need $c_n = 25\log(n)^{-\frac{3}{4}}$. For this value of c_n, we see that the limiting distribution is indeed a Gumbel distribution.

8. *An insurer models aggregate daily losses with a distribution in the MDA of a Fréchet distribution with $\xi = 0.8$. In the past 100 years, there have been 21 years including daily losses exceeding \$500,000, and 9 years including daily losses exceeding \$1,000,000. What is the probability of a daily loss exceeding \$2,000,000 during the next year? [10 mins]*

   We have that $\frac{M_{365} - d_{365}}{c_{365}}$ follows a Fréchet distribution with $\xi = 0.8$. The distribution function of this is $F(x) = e^{-(1+0.8x)^{-\frac{1}{0.8}}}$. We have that $P(M_{365} < 500000) = 0.79$ and $P(M_{365} < 1000000) = 0.91$. Solving $F(x) = 0.83$ gives

$x = \frac{(-\log(0.79))^{-0.8}-1}{0.8} = 2.72181880445$ and solving $F(x) = 0.91$ gives $x = \frac{(-\log(0.91))^{-0.8}-1}{0.8} = 7.0153512425$. Thus, we have $\frac{500000-d_{365}}{c_{365}} = 2.72181880445$ and $\frac{1000000-d_{365}}{c_{365}} = 7.0153512425$. We solve the equations

$$\begin{aligned} 2.72181880445c_{365} + d_{365} &= 500000 \\ 7.0153512425c_{365} + d_{365} &= 1000000 \\ 4.29353243805c_{365} &= 500000 \\ c_{365} &= 116454.226727 \\ d_{365} &= 183032.695837 \end{aligned}$$

Thus, the probability that the next year includes a daily loss exceeding \$2,000,000 is

$$\begin{aligned} P(M_{365} > 2000000) &= P\left(\frac{M_{365} - d_{365}}{c_{365}} > \frac{2000000 - 183032.695837}{116454.226727}\right) \\ &= 1 - F(15.6024161186) \\ &= 1 - e^{-(1+0.8\times 15.6024161186)^{-\frac{1}{0.8}}} \\ &= 0.037969196024 \end{aligned}$$

9. *A reinsurer offers an excess-of-loss reinsurance contract on a portfolio with attachment point \$10,000,000 and no policy limit. The aggregate loss distribution is estimated to lie in the MDA of a Gumbel distribution. The reinsurer estimates that the probability of paying a claim is 0.08 and the expected payment on the contract is \$4,800. What is the expected square of the payment on the contract. [10 mins]*

   Since the distribution is in the MDA of a Gumbel distribution, the excess loss distribution converges to an exponential distribution. The probability of a payment is 0.08 and the expected payment is \$4,800, so the conditional expected payment given that there is a payment is $\frac{4800}{0.08} = 60000$. The excess loss distribution is exponential with mean β, so we have $\beta = \$60,000$. The expected square of the payment conditional on a payment being made is $2 \times 60000^2 = 7200000000$. The expected square of the payment is therefore $0.08 \times 7200000000 = 576000000$.

10. *An insurer estimates that the time to completion of a claim comes from a distribution in the MDA of a GEV distribution with $\xi = -1.8$. The maximum time to completion is 20 years. They find that 2% of claims are incomplete after 5 years. Assuming the GPD approximation applies above 5 years, what percentage of claims are incomplete after 10 years?*

    Under the GPD approximation, the remaining time to completion after 5 years follows a GPD with $\xi = -1.8$. The maximum value is $-\frac{\beta}{\xi} = 15$, which gives $\beta = 1.8 \times 15 = 27$. The probability that a claim that is incomplete after 5 years remains incomplete 5 years later is $\left(1 - 1.8\frac{5}{27}\right)^{\frac{1}{1.8}} = 0.798309914099$. Thus, the percentage of claims that are incomplete after 10 years is $2 \times 0.798309914099 = 1.597\%$.

11. *An actuary is reviewing a sample of 483,230 observations that she believes comes from the MDA of a Fréchet distribution. She uses the Hill estimator to estimate ξ. She uses the $j = 481000$th order statistic as the threshold for the Hill estimator. Using this threshold, she gets the estimate $\xi = 1.45$. The order statistics near to this one are given in the following table:*

j	$x_{(j)}$
480999	*594303*
481000	*599045*
481001	*615667*
481002	*630520*
481003	*649402*
481004	*682034*
481005	*684215*
481006	*690144*

What value of ξ would she have calculated if she had used $j = 481005$? [15 mins]

The Hill estimator is given by

$$\hat{\xi} = \frac{1}{N-j+1} \sum_{k=j+1}^{N} \log(x_{(k)}) - \log(x_{(j)})$$

Thus, we have that

$$\frac{1}{2231} \sum_{k=j+1}^{N} \log(x_{(k)}) - \log(599045) = 1.45$$

This gives

$$\sum_{k=481001}^{483230} \log(x_{(k)}) = 2231(\log(599045) + 1.45) = 32914.1482509$$

We therefore have

$$\sum_{k=481006}^{483230} \log(x_{(k)}) = 32914.1482509 - \log(615667) - \log(630520) - \log(649402) - \log(682034) - \log(684215) = 32847.2108197$$

The Hill estimator using $j = 481005$ as the threshold is therefore

$$\frac{1}{2226} \sum_{k=481006}^{483230} \log(x_{(k)}) - \log(x_{(481005)}) = \frac{32847.2108197}{2226} - 13.4360274747 = 1.3201319232$$

12. *Loss amounts follow an exponential distribution with $\theta = 60,000$. The distribution of the number of losses is given in the following table:*

Number of Losses	*Probability*
0	*0.04*
1	*0.54*
2	*0.27*
3	*0.15*

Assume all losses are independent and independent of the number of losses. The insurance company buys excess-of-loss reinsurance on the part of the loss above \$150,000. Calculate the expected payment for this excess-of-loss reinsurance. [15 mins]

If the number of losses is n, then the aggregate loss follows a gamma distribution with $\alpha = n$ and $\theta = 60000$. The expected payment on the excess-of-loss insurance is therefore

$$\begin{aligned}
&\int_{150000}^{\infty} (x - 150000) \frac{x^{n-1} e^{-\frac{x}{60000}}}{(n-1)! 60000^n} \, dx \\
&= \int_{150000}^{\infty} \frac{x^n e^{-\frac{x}{60000}}}{(n-1)! 6000^n} \, dx - 150000 \int_{150000}^{\infty} \frac{x^{n-1} e^{-\frac{x}{60000}}}{(n-1)! 60000^n} \, dx \\
&= \int_{2.5}^{\infty} \frac{60000 n u^n e^{-u}}{n!} \, du - 150000 \int_{2.5}^{\infty} \frac{u^{n-1} e^{-u}}{(n-1)!} \, du
\end{aligned}$$

This gives the following expected payments on the excess-of-loss reinsurance:

Number of Losses	Probability	Expected payment on excess-of-loss	product
0	0.04	0	0
1	0.54	$60000 \times 1 \times 0.2872975 - 150000 \times 0.0820850 = 4925.10$	2659.554
2	0.27	$60000 \times 2 \times 0.5438131 - 150000 \times 0.2872975 = 22162.95$	5983.996
3	0.15	$60000 \times 3 \times 0.7575761 - 150000 \times 0.5438131 = 54791.74$	8218.760

The total expected payment on the excess-of-loss reinsurance is therefore $2659.554+5983.996+8218.760 = \$16,862.31$.

13. *Claim frequency follows a negative binomial distribution with $r = 5$ and $\beta = 2.9$. Claim severity (in thousands) has the following distribution:*

Severity	*Probability*
0	*0*
1	*0.600*
2	*0.220*
3	*0.166*

Use the recursive method to calculate the exact probability that aggregate claims are at least 4. [15 mins]

For the negative binomial distribution, we have $a = \frac{\beta}{1+\beta} = \frac{2.9}{3.9}$ and $b = \frac{(r-1)\beta}{1+\beta} = \frac{4 \times 2.9}{3.9}$, so the recursive formula

$$f_S(x) = \frac{(p_1 - (a+b)p_0) f_X(x) + \sum_{i=1}^{x} \left(a + \frac{bi}{x}\right) f_X(i) f_S(x-i)}{1 - a f_X(0)}$$

becomes

$$f_S(x) = \sum_{i=1}^{x} \frac{2.9}{3.9} \left(1 + \frac{4i}{x}\right) f_X(i) f_S(x-i)$$

Since the severity distribution has no probability at zero, the only way for the aggregate loss to be zero is if the frequency is zero, the probability of which is $\left(\frac{1}{1+\beta}\right)^r = \frac{1}{3.9}^5 = 0.00110835$. We now use the recurrence:

$$f_S(1) = \frac{2.9}{3.9} \times 5 \times 0.600 \times 0.00110835 = 0.002472473$$
$$f_S(2) = \frac{2.9}{3.9} \times (3 \times 0.600 \times 0.002472473 + 5 \times 0.220 \times 0.00110835) = 0.004215883$$
$$f_S(3) = \frac{2.9}{3.9} \times \left(\frac{7}{3} \times 0.600 \times 0.004215883 + \frac{11}{3} \times 0.220 \times 0.002472473 + 5 \times 0.166 \times 0.00110835\right) = 0.006555954$$

The probability that the aggregate payments exceed 4 is therefore $1 - 0.00110835 - 0.002472473 - 0.004215883 - 0.006555954 = 0.9856473$.

14. *Using an arithmetic distribution ($h = 1$) to approximate a Weibull distribution with $\tau = 3$ and $\theta = 12$, calculate the probability that the value is between 3.5 and 8.5, for the approximation using:*

(a) The method of rounding. [10 mins]

The method of rounding preserves this probability, since it assigns all values between 3.5 and 4.5 to 4, etc. Therefore this probability is $e^{-\left(\frac{3.5}{12}\right)^3} - e^{-\left(\frac{8.5}{12}\right)^3} = 0.2745978$.

(b) The method of local moment matching, matching 1 moment on each interval. [$\Gamma\left(\frac{4}{3}\right) = 0.8929795$.] [15 mins]

Using local moment matching, the probabilities of the intervals $[4, 5]$, $[5, 6]$, $[6, 7]$ and $[7, 8]$ are preserved, so the probability of these intervals is $e^{-\left(\frac{4}{12}\right)^3} - e^{-\left(\frac{8}{12}\right)^3} = 0.240929357799$

For the interval $[3, 4]$, the probability of this interval is $e^{-\left(\frac{3}{12}\right)^3} - e^{-\left(\frac{4}{12}\right)^3} = 0.020855992704$ while the conditional mean times this probability is

$$\begin{aligned}
\int_3^4 x \left(\frac{3x^2}{12^3} e^{-\left(\frac{x}{12}\right)^3}\right) dx &= \int_{\left(\frac{3}{12}\right)^3}^{\left(\frac{4}{12}\right)^3} 12\sqrt[3]{u} e^{-u}\, du \\
&= 12 \int_{\frac{1}{64}}^{\frac{1}{27}} u^{\frac{1}{3}} e^{-u}\, du \\
&= 12\Gamma\left(\frac{4}{3}\right)\left(S_{\text{Gamma}}\left(\frac{1}{64}, \alpha = \frac{4}{3}\right) - S_{\text{Gamma}}\left(\frac{1}{27}, \alpha = \frac{4}{3}\right)\right) \\
&= 0.07394568
\end{aligned}$$

We are now trying to solve for p_3 and p_4 such that

$$\begin{aligned}
p_3 + p_4 &= 0.020855992704 \\
4p_3 + 4p_4 &= 0.07394568 \\
p_4 &= 0.07394568 - 3 \times 0.020855992704 = 0.011377701888
\end{aligned}$$

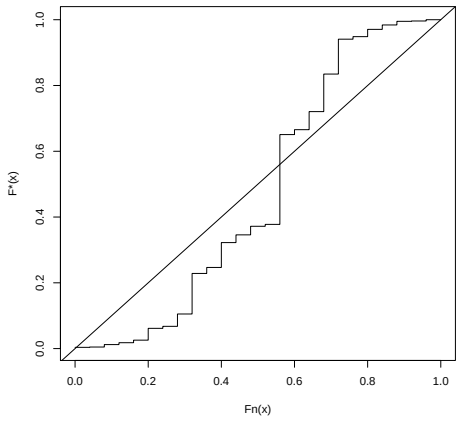
For the interval $[8, 9]$, the probability of this interval is $e^{-\left(\frac{8}{12}\right)^3} - e^{-\left(\frac{9}{12}\right)^3} = 0.087751067934$, while the conditional mean times this probability is

$$12\Gamma\left(\frac{4}{3}\right)\left(S_{\text{Gamma}}\left(\frac{8}{27}, \alpha = \frac{4}{3}\right) - S_{\text{Gamma}}\left(\frac{27}{64}, \alpha = \frac{4}{3}\right)\right) = 0.7466863$$

$$\begin{aligned} p_8 + p_9 &= 0.087751067934 \\ 8p_8 + 9p_9 &= 0.7466863 \\ p_8 &= 9 \times 0.087751067934 - 0.7466863 = 0.043073311406 \end{aligned}$$

So the probability of the interval $[3.5, 8.5]$ is therefore $0.240929357799 + 0.011377701888 + 0.043073311406 = 0.295380371093$.

15. *An insurance company collects a sample of 25 past claims, and attempts to fit a Pareto distribution to the claims. Based on experience with other claims, the company believes that a Pareto distribution with $\alpha = 3.5$ and $\theta = 4,600$ may be appropriate to model these claims. It constructs the following p-p plot to compare the sample to this distribution:*

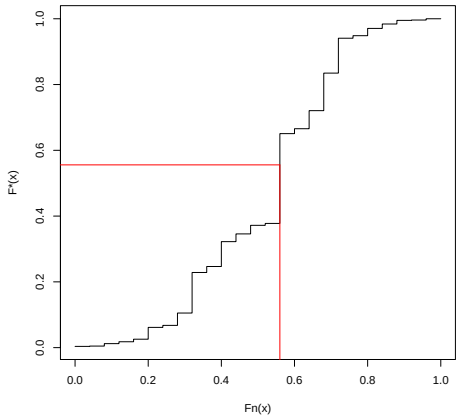


(a) How many of the points in their sample were less than 1,200? [5 mins.]

We have

$$F^*(1200) = 1 - \left(\frac{46}{58}\right)^{3.5} = 0.5557224$$

so we look for the point on the graph with $F^*(x) = 0.5557224$.



We see that the corresponding value of $F_n(x)$ is 0.56. (The values of $F_n(x)$ are in increments of 0.04, since there are 25 data points. The value corresponding to $F^*(x)$ is one increment before 0.6, so is 0.56).

(b) Which of the following statements best describes the fit of the Pareto distribution to the data: [5 mins.]

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

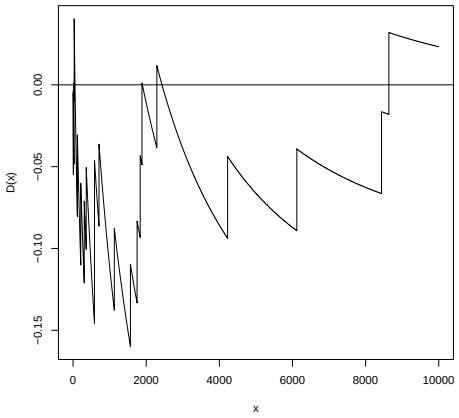
(ii) The Pareto distribution assigns too much probability to low values and too little probability to high values.

(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 there are 8 data points with $F^*(x) < 0.1$ approximately. The expected number is 2.5. There are 7 data points with $F^(x) > 0.9$. Again, the expected number is 2.5. The Pareto distribution has therefore underestimated the probabilities of these tail regions, and overestimated the probability of the region in between. Therefore, statement (iv) best describes the fit.

16. *An insurance company collects a sample of 20 claims. Based on previous experience, it believes these claims might follow a Weibull distribution with $\tau = 0.6$ and a known value of θ. To test this, it obtains a plot of $D(x)$.*



(a) Which of the following is the value of θ used in the plot: [5 mins.]

(i) 800

(ii) 1,100

(iii) 2,200

(iv) 3,500

The data points in the sample correspond to vertical line segments on the plot. We see for example, that there are 3 data points above 6000, so $F_{20}(6000) = \frac{17}{20} = 0.85$. Reading from the graph, we get that $D(6000) \approx -0.09$. This means $F^*(6000) = 0.85 - (-0.09) = 0.94$. This gives:

$$\begin{aligned} 1 - e^{-\left(\frac{6000}{\theta}\right)^{0.6}} &= 0.94 \\ \left(\frac{6000}{\theta}\right)^{0.6} &= -\log(0.06) \\ \frac{6000}{\theta} &= (-\log(0.06))^{\frac{1}{0.6}} \\ \theta = \frac{6000}{(-\log(0.06))^{\frac{1}{0.6}}} &= 1070.112 \end{aligned}$$

This is clearly closest to (ii), so (ii) is the value of θ used. (The difference between this answer and the 1,100 is because we only have limited accuracy reading the graph.)

[We can find the value of θ by reading off the value of $D(x)$ for any X on the graph. If it is difficult to count the number of vertical line segments, we could compare $D(x_1)$ and $D(x_2)$ for values of x_1 and x_2 with no vertical line segments in between. For example, we can read the value $D(4200) \approx -0.04$, which leads us to solve

$$F^*(6000) - F^*(4200) = 0.05$$

We can try the values given to see which is closer to the solution.]

(b) Which of the following statements best describes the fit of the Weibull distribution to the data: [5 mins.]

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

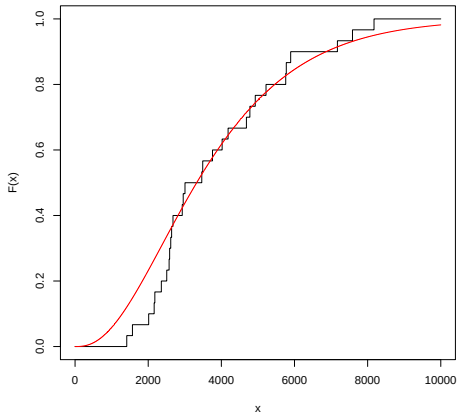
(ii) The Weibull distribution assigns too much probability to low values and too little probability to high values.

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

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

Recall that $D(x) = F_n(x) - F^*(x)$, so if $D(x) < 0$, we have $F^*(x) > F_n(x)$, while if $D(x) > 0$, we have $F^*(x) < F_n(x)$. On the graph shown, we have that $D(x)$ is nearly always negative for the range of the data. [Technically, it is positive for all values larger than the data sample, but this always happens, because for the largest value of the data sample, we have $F_n(x) = 1 > F^*(x)$.] This means that $F^*(x) > F_n(x)$ for most x in the range. This means that the Weibull distribution assigns more probability to smaller values of x, and less probability to larger values of x, which is statement (ii).

17. *An insurance company collects a sample of 30 claims. Based on previous experience, it believes these claims might follow a gamma distribution with $\alpha = 2.7$ and $\theta = 1400$. To test this, it compares plots of $F_n(x)$ and $F_*(x)$.*



(a) Which of the following is the value of the Kolmogorov-Smirnov statistic for this model and this data [5 mins.]

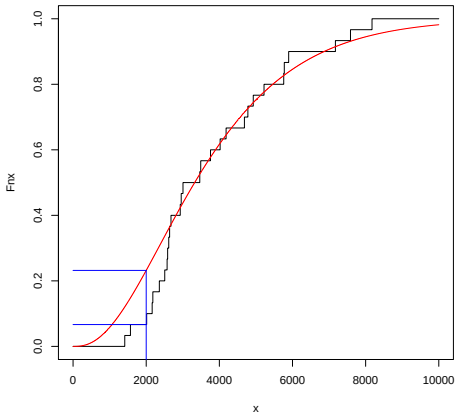
(i) 0.0102432

(ii) 0.0450353

(iii) 0.0924252

(iv) 0.1678255

The Kolmogorov-Smirnov test statistic is the maximum value of the absolute difference between the empirical and model distribution functions, that is $|F_n(x) - F^*(x)|$. On the graph, we see this happens at around 2000, and read the values from the graph:



We read $F_n(x) = 0.066667$ (we know the possible values of $F_n(x)$, since we know there are 30 data points), and $F^*(x) = 0.23$ (the actual value is 0.2318889.) The difference is therefore about 1.6, so (iv) is the correct answer.

(b) Which of the following statements best describes the fit of the Gamma distribution to the data: [5 mins.]

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

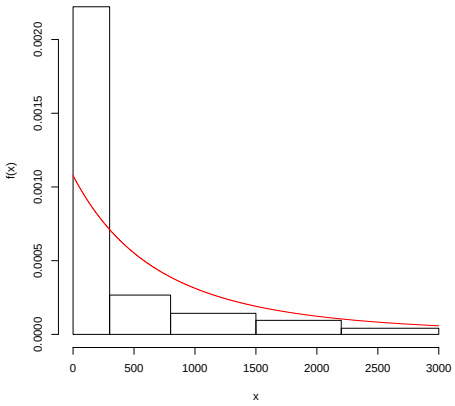
(ii) The Gamma distribution assigns too much probability to low values and too little probability to high values.

(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.

From the graph, we see that $F*(x)$ is too large for small values less than about 2500, and about correct for larger values. This means that the gamma model assigns too little probability in the range 0–2,000 and too much in the range 2,000–2,500. We also see that $F^*(x)$ is slightly too low at values above 6,000. This means that the gamma distribution assigns too little probability to values larger than 6,000. This means that (iv) is probably the best description of the fit. However, a case could be made for (ii) being a good description, since the difference between $F^*(x)$ and $F_n(x)$ for $x > 6000$ is very small.

18. *An insurance company collects a sample of 30 past claims, and attempts to fit a Pareto distribution to the claims. Based on experience with other claims, the company believes that a Pareto distribution with $\alpha = 2.8$ and $\theta = 2,600$ may be appropriate to model these claims. It compares the density functions in the following plot:*

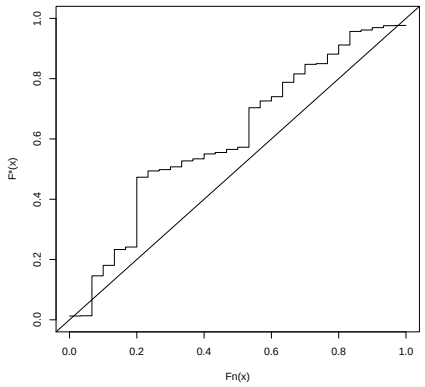


(a) How many data points in the sample were between 1500 and 3000? [5 mins.]

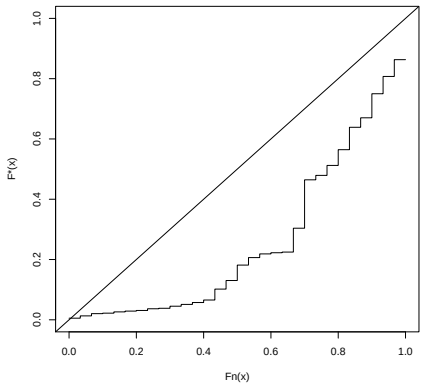
We are asking how many data points are in the last two bars. The height of the fourth bar (from 1,500–2,200) is about 0.0001, and the height of the fifth bar (from 2,200–3,000) is about 0.00005, so the areas of these two bars are $700 \times 0.0001 = 0.07$ and $800 \times 0.00005 = 0.04$ respectively. Since there are 30 claims in the sample, these correspond to 2 data points and 1 data point respectively, (which would give accurate heights of 0.00009524 and 0.00004167 respectively). Therefore, the number of data points between 1,500 and 3,000 is 3.

(b) Which of the following plots is the p-p plot for this data and model? [10 mins.]

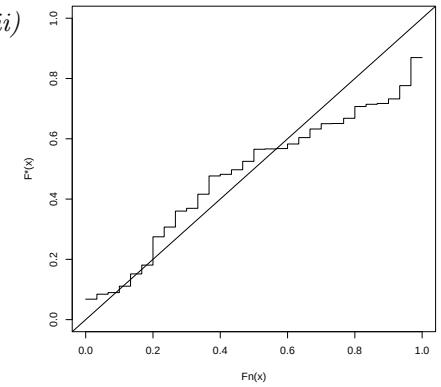
(i)



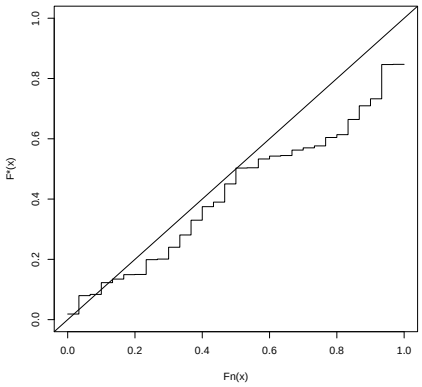
(ii)



(iii)



(iv)



[15 mins]

From the histogram, we see that the model assigns too little probability to small values less than 300, and too much probability to values more than 500. The p-p plot should therefore have slope less than 1 for the first part, then slope more than 1. We would expect $F_n(x) > F^*(x)$ for all x, so the p-p plot should be entirely below the line $y = x$ (it is in theory possible there could be some small values with $F*(x) > F_n(x)$, since the histogram only shows grouped data, so it is possible for example that all samples in the range 0–300 actually fell in the range 200–300). It seems that the largest difference between $F_n(x)$ and $F^*(x)$ should happen at around $x = 500$, and it looks like the area of the bar 0–300 on the histogram is approximately equal to the combined area of the other 3 bars. More accurately, it looks like the height of this bar is about 0.0022, and the width is about 300, so the area is about 0.66, so the largest difference between $F_n(x)$ and $F^*(x)$ should occur at about $F_n(x) = 0.66$. Also, after the first bar, the model is overestimating the probability density, which means that after this point, the slope of the p-p plot should be more than 1.

Looking at the options, plots (i) and (iii) are above the $y = x$ line for some values of x. Plot (iv) is close to the line for values less than $F_n(x) = 0.5$, and does not deviate so much from the line, and its furthest point from the line is around $F_n(x) = 0.9$, so it is not correct. Therefore, plot (ii) is the correct plot.

19.

20. *An insurance company collects the following sample:*

    `2.31 8.65 35.29 42.27 151.51 194.99 523.50 1262.01 1402.72 6063.74`

    They model this as following a Pareto distribution with $\alpha = 2$ *and* $\theta = 2000$. *Calculate the Kolmogorov-Smirnov statistic for this model and this data. [10 mins.]*

x	$F*(x)$	$D(x^+)$	$D(x^-)$	
2.31	0.002306004	0.002306004	0.09769400	0.09769400
8.65	0.008594205	0.091405795	0.19140580	0.19140580
35.29	0.034377462	0.165622538	0.26562254	0.26562254
42.27	0.040966725	0.259033275	0.35903327	0.35903327
151.51	0.135881599	0.264118401	0.36411840	0.36411840
194.99	0.169776735	0.330223265	0.43022327	0.43022327
523.50	0.371864450	0.228135550	0.32813555	0.32813555
1262.01	0.624085208	0.075914792	0.17591479	0.17591479
1402.72	0.654532208	0.145467792	0.24546779	0.24546779
6063.74	0.938484160	0.038484160	0.06151584	0.06151584

So the Kolmogorov-Smirnov statistic is 0.4302.

21. *An insurance company collects the following sample:*

    0.27 2.03 9.89 16.96 28.38 236.46 268.36 453.19 633.26 718.68 1414.59 1588.19 2535.69 4937.93 5431.13

    They model this as following a gamma distribution with $\alpha = 0.4$ and $\theta = 6000$. Calculate the Anderson-Darling statistic for this model and this data. [10 mins.]

    You are given the following values of the Gamma distribution used in the model:

x	$F(x)$	$\log(F(x))$	$\log(1-F(x))$
0.27	*0.02056964*	−3.8839392	−0.02078414
2.03	*0.04609387*	−3.0770753	−0.04719001
9.89	*0.08680820*	−2.4440542	−0.09080935
16.96	*0.10767291*	−2.2286572	−0.11392253
28.38	*0.13222244*	−2.0232696	−0.14181987
236.46	*0.30572308*	−1.1850755	−0.36488438
268.36	*0.32111513*	−1.1359556	−0.38730373
453.19	*0.39258278*	−0.9350079	−0.49853938
633.26	*0.44506880*	−0.8095264	−0.58891114
718.68	*0.46633756*	−0.7628455	−0.62799177
1414.59	*0.59250242*	−0.5234003	−0.89772028
1588.19	*0.61583950*	−0.4847689	−0.95669484
2535.69	*0.71295893*	−0.3383315	−1.24812996
4937.93	*0.84646394*	−0.1666877	−1.87381984
5431.13	*0.86352967*	−0.1467270	−1.99164807

The Anderson-Darling statistic for complete data with no truncation or censorship can be calculated as

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

We compute the terms in the following table:

y_j	$n(1-F_n(y_j))^2\left(\log(1-F^*(y_j))-\log(1-F^*(y_{j+1}))\right)$	$n(F_n(y_j))^2\left(\log(F^*(y_{j+1}))-\log(F^*(y_j))\right)$
0.00	0.311762095	
0.27	0.345036701333	0.0537909266667
2.03	0.491444564001	0.1688056266667
9.89	0.221886528	0.1292382000000
16.96	0.225038542667	0.2190801066680
28.38	1.48709673333	1.3969901666700
236.46	0.12106449	0.1178877600000
268.36	0.474605439999	0.6564291533340
453.19	0.295214416001	0.5353877333330
533.26	0.093793512	0.2520768600000
718.68	0.449547516666	1.5963013333400
414.59	0.0629061973335	0.3116266266660
588.19	0.174861072	1.4057990400000
535.69	0.166850634666	1.9338534800000
937.93	0.00785521533335	0.2608198133330
431.13	2.2009050000000	1.917233
	4.92896366333	11.2389918267

This gives $A^2 = 4.928964 + 11.2389918267 - 15 = 1.1679558267$.

22. *An insurance company collects the following sample:*

105.13 304.10 323.11 359.09 360.43 368.63 413.47 448.81 606.88 612.58 930.35 1002.37 1161.78 1205.25 5585.37

They want to decide whether this data is better modeled as following an inverse gamma distribution, or an inverse exponential distribution. They calculate that the MLEs for the inverse gamma distribution as $\alpha = 1.695545$ *and* $\theta = 705.7664$*, and the MLE for the inverse exponential distribution as* $\theta = 416.2476$*. They also calculate, for this data that* $\sum_{i=1}^{15}\log(x_i) = 95.31415$ *and* $\sum_{i=1}^{15}\frac{1}{x_i} = 0.03603625$*, and that* $\Gamma(1.695545) = 0.9078021$*. You are given the following table of critical values for the chi-squared distribution at the 5% significance level. Indicate in your answer which critical value you are using. [15 mins.]*

Degrees of Freedom	*95% critical value*
1	*3.841459*
2	*5.991465*
3	*7.814728*
4	*9.487729*
5	*11.070498*

For the inverse gamma distribution, the log-likelihood of the data point x is

$$\log\left(\frac{705.7664^{1.695545}e^{-\frac{705.7664}{x}}}{x^{2.695545}\Gamma(1.695545)}\right) = 1.695545\log(705.7664) - \log(\Gamma(1.695545)) - 2.695545\log(x) - \frac{705.7664}{x}$$
$$= 11.21829 - 2.695545\log(x) - \frac{705.7664}{x}$$

The total log-likelihood of the data is therefore

$$\begin{aligned}11.21829 \times 15 - 2.695545\,(&\log(105.13) + \log(304.10) + \log(323.11) + \log(359.09) + \log(360.43) + \log(368.63) + \log(413.47) + \\ &\log(448.81) + \log(606.88) + \log(612.58) + \log(930.35) + \log(1002.37) + \log(1161.78) + \log(1205.25) + \log(5585.37)) \\ -705.7664\left(\frac{1}{105.13} + \frac{1}{304.10} + \frac{1}{323.11} + \frac{1}{359.09} + \frac{1}{360.43} + \frac{1}{368.63} + \frac{1}{413.47} + \frac{1}{448.81} + \frac{1}{606.88} + \frac{1}{612.58} + \frac{1}{930.35} + \right.&\\ \left. \frac{1}{1002.37} + \frac{1}{1161.78} + \frac{1}{1205.25} + \frac{1}{5585.37}\right)&\\ = -114.0824&\end{aligned}$$

For the inverse exponential, the log-likelihood of the data point x is

$$\log\left(\frac{416.2476}{x^2}e^{-\frac{416.2476}{x}}\right) = 6.03128 - 2\log(x) - \frac{416.2476}{x}$$

The log-likelihood of the data is therefore

$$\begin{aligned}6.03128 \times 15 - 2\,(&\log(105.13) + \log(304.10) + \log(323.11) + \log(359.09) + \log(360.43) + \log(368.63) + \log(413.47) + \\ &\log(448.81) + \log(606.88) + \log(612.58) + \log(930.35) + \log(1002.37) + \log(1161.78) + \log(1205.25) + \log(5585.37)) \\ -416.2476\left(\frac{1}{105.13} + \frac{1}{304.10} + \frac{1}{323.11} + \frac{1}{359.09} + \frac{1}{360.43} + \frac{1}{368.63} + \frac{1}{413.47} + \frac{1}{448.81} + \frac{1}{606.88} + \frac{1}{612.58} + \frac{1}{930.35} + \right.&\\ \left. \frac{1}{1002.37} + \frac{1}{1161.78} + \frac{1}{1205.25} + \frac{1}{5585.37}\right)&\\ = -115.1591&\end{aligned}$$

The likelihood ratio statistic is therefore $2(-114.0824 - (-115.1591)) = 2.1534$. This should be compared to the chi-square distribution with one degree of freedom (since the inverse gamma has 2 degrees of freedom, and the inverse exponential has 1). The critical value for this is 3.841459, so the statistic is not significant. This means there is not sufficient evidence that the inverse gamma distribution fits the data better.

23. *An insurance company collects the following sample:*

    `0.1 0.2 0.3 2.1 16.8 28.4 45.7 53.5 74.2 99.5 159.3 183.5 206.3 273.9 461.9 482.9 1118.5 1444.7 2084.3 3984.8`

    They want to decide whether this data is better modeled as following an inverse exponential distribution or a Weibull distribution. They calculate that the MLE for the inverse exponential distribution is $\theta = 1.052901$*, and the corresponding likelihood is* -183.51*. They also calculate that for the Weibull distribution, the MLE is* $\tau = 0.48$*,* $\theta = 255.2235$*. The log-likelihood is therefore* -141.8325*. Use AIC and BIC to determine which distribution is a better fit for the data. [5 mins.]*

    The AIC is $l(x) - p$, while the BIC is $l(x) - \frac{p}{2}\log(n)$. For the inverse exponential distribution, we have $p = 1$, while for the Weibull distribution, we have $p = 2$. For this data set, we have $n = 20$, so the AIC and BIC are:

odel	AIC	BIC
verse Exponential	$-183.51 - 1 = -184.51$	$-183.51 - \frac{1}{2}\log(20) = -185.007866137$
eibull	$-141.8325 - 2 = -143.8325$	$-141.8325 - \frac{2}{2}\ln(20) = -144.828232274$

Therefore, both AIC and BIC prefer the Weibull distribution.

24. *An insurance company collects the following data sample on claims data*

Claim Amount	*Number of Claims*
Less than \$5,000	*1,026*
\$5,000–\$10,000	*850*
\$10,000–\$20,000	*1,182*
\$20,000–\$50,000	*942*
More than \$50,000	*573*

Its previous experience suggests that the distribution should be modelled as following a Pareto distribution with $\alpha = 3$ *and* $\theta = 28,000$*. Perform a chi-squared test to determine whether this distribution is a good fit for the data at the 95% level. [10 mins.]*

You may use the following critical values for the chi-squared distribution:

Degrees of Freedom	*95% critical value*
1	*3.841459*
2	*5.991465*
3	*7.814728*
4	*9.487729*
5	*11.070498*

The expected frequencies of each interval are:

$$4573\left(1-\left(\frac{28}{33}\right)^3\right) = 1779.598$$
$$4573\left(\left(\frac{28}{33}\right)^3-\left(\frac{28}{38}\right)^3\right) = 963.9355$$
$$4573\left(\left(\frac{28}{38}\right)^3-\left(\frac{28}{48}\right)^3\right) = 921.7474$$
$$4573\left(\left(\frac{28}{48}\right)^3-\left(\frac{28}{78}\right)^3\right) = 696.1798$$
$$4573\left(\frac{28}{78}\right)^3 = 211.5395$$

Therefore, the chi-squared statistic is

$$\frac{(1026-1779.598)^2}{1779.598}+\frac{(850-963.9355)^2}{963.9355}+\frac{(1182-921.7474)^2}{921.7474}+\frac{(942-696.1798)^2}{696.1798}+\frac{(573-211.5395)^2}{211.5395} = 1110.503$$

Since the parameters are not estimated the number of degrees of freedom is $5-1=4$, so the critical value is 9.487729. The null hypothesis is rejected. The data do not fit the model well.