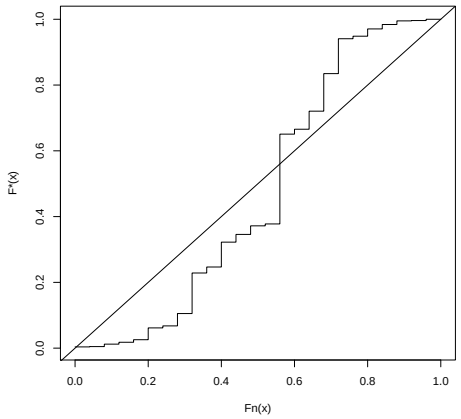
ACSC/STAT 4703, Actuarial Models II
Fall 2015
Toby Kenney
Sample Final Examination
Model Solutions

For each question that asks you to simulate a small number of samples from a distribution, use the following simulated uniform values, starting from the first, and using as many numbers as needed for the question. Go back to the first value at the start of each part question.

```
0.58665797 0.12487271 0.87530540 0.49197147 0.55262301 0.14644543 0.89151074 0.46559276 0.42856173
0.63507522 0.78161985 0.69613284 0.37786683 0.51447243 0.48952100 0.28195163 0.62179048 0.66186936
0.42715830 0.70003263 0.59328856 0.97308150 0.14087141 0.08049598 0.98662077 0.91974635 0.56037580
0.07804151 0.48363702 0.33763780
```

1. *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:*

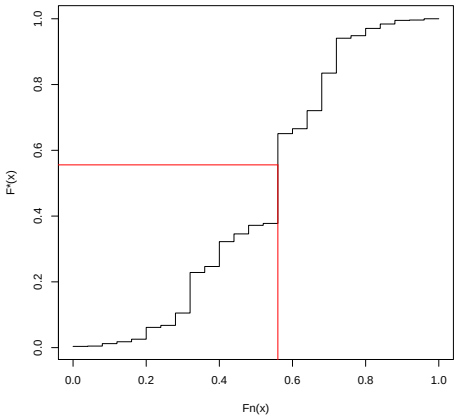


(a) How many of the points in their sample were less than 1,200?

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:

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

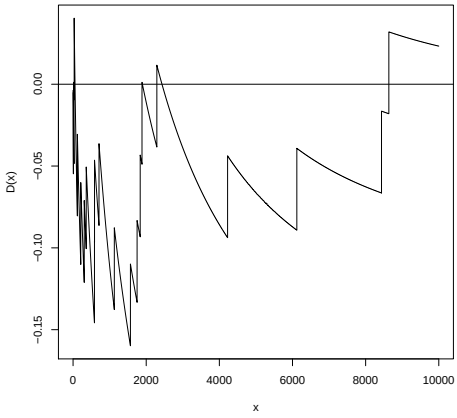
(ii) The Pareto distribution assigns too much probability to low values and too little probability to high values.

(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.

2. *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:

(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:

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

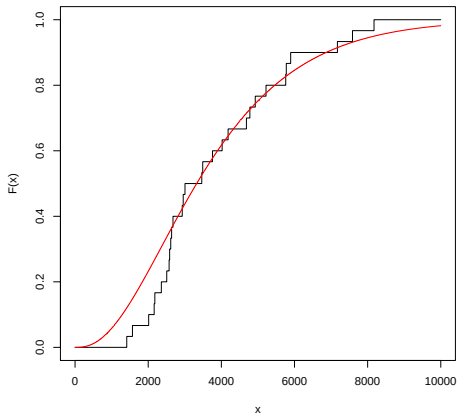
(ii) The Weibull distribution assigns too much probability to low values and too little probability to high values.

(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).

3. *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

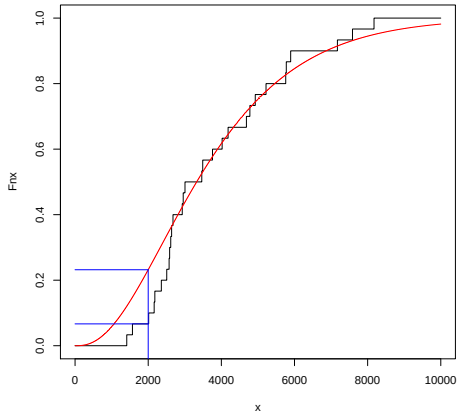
(i) 0.0102432

(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:

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

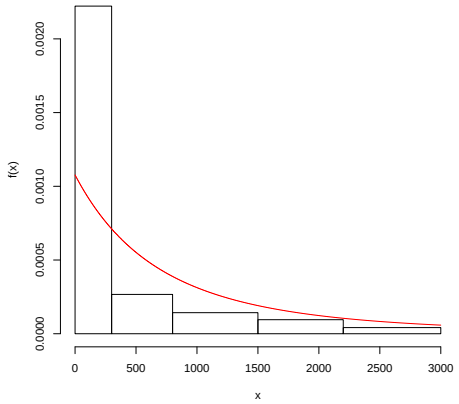
(ii) The Gamma distribution assigns too much probability to low values and too little probability to high values.

(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.

4. *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:*

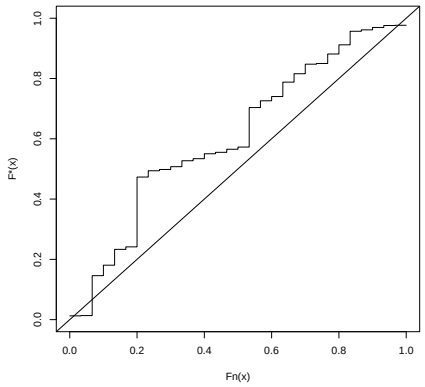


(a) How many data points in the sample were between 1500 and 3000?

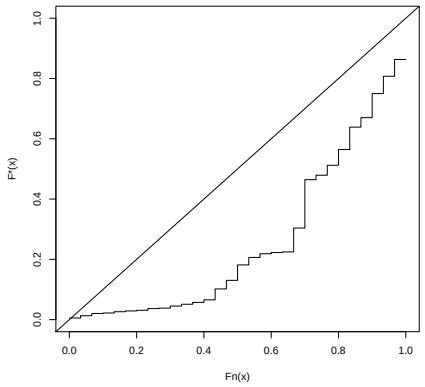
We are asking how many data points are in the last two bars. The height of the fourth bar (from 1,500–2,200) is about 0.0001, and the height of the fifth bar (from 2,200–3,000) is about 0.00005, so the areas of these two bars are $700 \times 0.0001 = 0.07$ and $800 \times 0.00005 = 0.04$ respectively. 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?

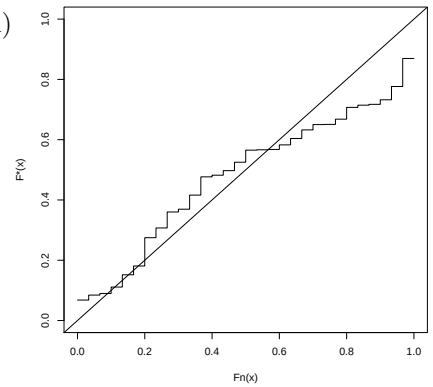
(i)



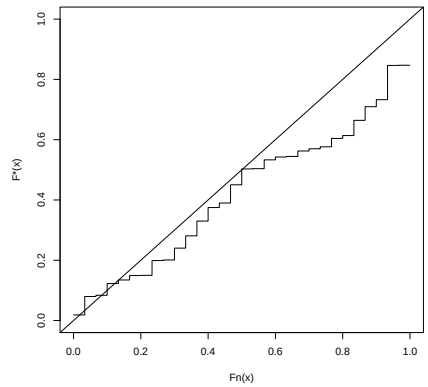
(ii)



(iii)



(iv)



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.

5. *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.

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.

6. *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.

   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:

j	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	0.00	0.311762095	
1	0.27	0.345036675	0.3649728
2	2.03	0.491444551	0.6910801
3	9.89	0.221886527	0.9992218
4	16.96	0.225038563	1.4317238
5	28.38	1.487096754	2.0290460
6	236.46	0.121064462	1.3379735
7	268.36	0.474605439	1.8103976
8	453.19	0.295214434	1.7395509
9	633.26	0.093793516	1.6241571
10	718.68	0.449547513	2.4527362
11	1414.59	0.062906196	1.2514677
12	1588.19	0.174861070	2.0328028
13	2535.69	0.166850636	2.2157161
14	4937.93	0.007855215	0.4465469
15	5431.13		1.917233
total	4.928964	22.34463	

This gives $A^2 = 4.928964 + 22.34463 - 15 = 12.27359$.

7. *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.*

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

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

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

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

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.

8. *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.*

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.

9. *An insurance company sells home insurance. It estimates that the standard deviation of the aggregate annual claim is \$5,326 and the mean is \$1,804.*

(a) How many years history are needed for an individual or group to be assigned full credibility? (Use $r = 0.05$, $p = 0.95$.)

The variance of the mean of a sample of n observations from an individual is $\frac{5326^2}{n}$, so a 95% confidence interval for this individual is their mean plus or minus $1.96 \times \frac{5326}{\sqrt{n}}$. We want the relative error in this estimate to be at most 5%. That is we want

$$1.96 \times \frac{5326}{\sqrt{n}} = 0.05 \times 1804$$
$$n = \left(\frac{10438.96}{90.2}\right)^2 = 13393.73$$

(b) What is the Credibility premium, using limited fluctuation credibility, for an individual who has claimed a total of \$42,381 in the past 19 years?

This individual's average annual aggregate claims are $\frac{42381}{19} = \$2230.58$. The credibility is $\sqrt{\frac{19}{13393.73}} = 0.03766396$, so the credibility premium is $0.03766396 \times 2230.58 + 0.96233604 \times 1804 = \$1820.07$.

10. *For a car insurance policy, the book premium for claim severity is \$2,300. An individual has made 7 claims in the past 12 years, with average claim severity \$1,074. Calculate the credibility estimate for claim severity for this individual using limited fluctuation credibility, if the standard for full credibility is:*

(a) 157 claims.

If the standard for full credibility is 157 claims, then this individual's credibility is $\sqrt{\frac{7}{157}} = 0.2111539$, and the credibility estimate is $0.2111539 \times 1074 + 0.7888461 \times 2300 = 2041.13$.

(b) 284 years.

If the standard for full credibility is 157 claims, then this individual's credibility is $\sqrt{\frac{12}{284}} = 0.2055566$, and the credibility estimate is $0.2055566 \times 1074 + 0.7944434 \times 2300 = 2047.99$.

11. *A worker's compensation insurance company classifies workplaces as "safe" or "hazardous". Claims from hazardous workplaces follow a Gamma distribution with $\alpha = 0.1021749$, $\theta = 1066798$ (mean \$109,000 and standard deviation \$341,000). Claims from safe workplaces follow a Gamma distribution with $\alpha = 0.01209244$, $\theta = 2646281$ (mean \$32,000 and standard deviation \$261,000). 94% of workplaces are classified as safe.*

[You may need the following values:

$$\Gamma(0.01209244) = 82.13091$$
$$\Gamma(0.1021749) = 9.302457$$

(a) Calculate the expectation and variance of claim size for a claim from a randomly chosen workplace.

The expectation is $0.94 \times 32000 + 0.06 \times 109000 = \$36,620$. The variance is $(109000 - 32000)^2 \times 0.94 \times 0.06 + 0.94 \times 261000^2 + 0.06 \times 341000^2 = 71,345,000,000$.

(b) The last 2 claims from a particular workplace are \$488,200 and \$17,400. Calculate the expectation and variance for the next claim size from this workplace.

If the workplace is safe, the likelihood of these claim sizes is

$$\left(\frac{488200^{-0.98790756}e^{-\frac{488200}{2646281}}}{2646281^{0.01209244}\Gamma(0.01209244)}\right)\left(\frac{17400^{-0.98790756}e^{-\frac{17400}{2646281}}}{2646281^{0.01209244}\Gamma(0.01209244)}\right) = 1.32923 \times 10^{-14}$$

If the workplace is hazardous, the likelihood of these claim sizes is

$$\left(\frac{488200^{-0.8978251}e^{-\frac{488200}{1066798}}}{1066798^{0.1021749}\Gamma(0.1021749)}\right)\left(\frac{17400^{-0.8978251}e^{-\frac{17400}{1066798}}}{1066798^{0.1021749}\Gamma(0.1021749)}\right) = 5.134517 \times 10^{-13}$$

The posterior probability that the workplace is safe is therefore $\frac{0.94\times 1.32923\times 10^{-14}}{0.94\times 1.32923\times 10^{-14}+0.06\times 5.134517\times 10^{-13}} = 0.2885502$, so the expectation is $0.2885502 \times 32000 + 0.7114498 \times 109000 = \$86,781.63$.

The variance is $77000^2 \times 0.2885502 \times 0.7114498 + 0.2885502 \times 261000^2 + 0.7114498 \times 341000^2 = 103,601,580,743$.

12. *An insurance company sets the book pure premium for its home insurance at \$791. The expected process variance is 6,362,000 and the variance of hypothetical means is 341,200. If an individual has no claims over the last 8 years, calculate the credibility premium for this individual's next year's insurance using the Bühlmann model.*

The credibility is $Z = \frac{8}{8+\frac{6362000}{341200}} = 0.3002332$. Therefore the premium is $0.6997668 \times 791 = \$553.52$.

13. *An insurance company is reviewing the premium for an individual with the following past claim history:*

Year	*1*	*2*	*3*	*4*	*5*
Exposure	*0.2*	*1*	*1*	*0.4*	*0.8*
Aggregate claims	*0*	*\$2,592*	*0*	*\$147*	*\$1,320*

The usual premium per unit of exposure is \$2,700. The expected process variance is 123045 and the variance of hypothetical means is 36403 (both per unit of exposure). Calculate the credibility premium for this individual if she has 0.6 units of exposure in year 6.

The credibility of the policyholder's experience is $\frac{3.4}{3.4+\frac{123045}{36403}} = 0.5014691$. The policyholder's aggregate claims were \$4,059, so average claims per unit of exposure are $\frac{4059}{3.4} = \$1,193.53$. The credibility premium per unit of exposure is therefore $0.5014691 \times 1193.53 + 0.4985309 \times 2700 = \$1,944.70$. This is for a whole unit of exposure. Since the policyholder has 0.6 units of exposure, the credibility premium is $0.6 \times 1944.70 = \$1,166.82$.

14. *An insurance company has 3 years of past history on a homeowner, denoted X_1, X_2, X_3. Because the individual moved house at the end of the second year, the third year has a different mean and variance, and is not as correlated with the other two years. It has the following*

$$\begin{aligned}
\mathbb{E}(X_1) &= 1,322 & \operatorname{Var}(X_1) &= 226,000 \\
\mathbb{E}(X_2) &= 1,322 & \operatorname{Var}(X_2) &= 226,000 \\
\mathbb{E}(X_3) &= 4,081 & \operatorname{Var}(X_3) &= 1,108,000 \\
\mathbb{E}(X_4) &= 4,081 & \operatorname{Var}(X_4) &= 1,108,000 \\
\operatorname{Cov}(X_1, X_2) &= 214 & \operatorname{Cov}(X_1, X_3) &= 181 \\
\operatorname{Cov}(X_2, X_3) &= 181 & \operatorname{Cov}(X_1, X_4) &= 181 \\
\operatorname{Cov}(X_2, X_4) &= 181 & \operatorname{Cov}(X_3, X_4) &= 861
\end{aligned}$$

It uses a formula $\hat{X_4} = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3$ *to calculate the credibility premium in the fourth year. Calculate the values of* α_0, α_1, α_2 *and* α_3.

The company needs to choose α_0, α_1, α_2 and α_3 to satisfy:

$$\begin{aligned}
\mathbb{E}(X_4) &= \alpha_0 + \alpha_1 \mathbb{E}(X_1) + \alpha_2 \mathbb{E}(X_2) + \alpha_3 \mathbb{E}(X_3) \\
\operatorname{Cov}(X_4, X_1) &= \alpha_1 \operatorname{Var}(X_1) + \alpha_2 \operatorname{Cov}(X_2, X_1) + \alpha_3 \operatorname{Cov}(X_3, X_1) \\
\operatorname{Cov}(X_4, X_2) &= \alpha_1 \operatorname{Cov}(X_1, X_2) + \alpha_2 \operatorname{Var}(X_2) + \alpha_3 \operatorname{Cov}(X_3, X_2) \\
\operatorname{Cov}(X_4, X_1) &= \alpha_1 \operatorname{Cov}(X_1, X_3) + \alpha_2 \operatorname{Cov}(X_2, X_3) + \alpha_3 \operatorname{Var}(X_3)
\end{aligned}$$

Substituting the values gives:

$$\begin{aligned}
4081 &= \alpha_0 + 1322\alpha_1 + 1322\alpha_2 + 4081\alpha_3 \\
181 &= 226000\alpha_1 + 214\alpha_2 + 181\alpha_3 \\
181 &= 214\alpha_1 + 226000\alpha_2 + 181\alpha_3 \\
861 &= 181\alpha_1 + 181\alpha_2 + 1108000\alpha_3
\end{aligned}$$

By symmetry, we see that α_1 and α_2 are equal. This gives

$$\begin{aligned}
181 &= 226214\alpha_1 + 181\alpha_3 \\
861 &= 362\alpha_1 + 1108000\alpha_3 \\
226214 \times 861 - 362 \times 181 &= (226214 \times 1108000 + 362 \times 181)\alpha_3 \\
\alpha_3 &= \frac{194704732}{250,645,046,478} = 0.0007768146 \\
\alpha_1 &= \frac{181 - 181 \times 0.0007768146}{226214} = 0.0007995058 \\
\alpha_0 &= 4081 - 1322 \times 2 \times 0.0007995058 - 4081 \times 0.0007768146 = 4075.716
\end{aligned}$$

The values are:

$$\alpha_0 = 4075.716$$
$$\alpha_1 = 0.0007995058$$
$$\alpha_2 = 0.0007995058$$
$$\alpha_3 = 0.0007768146$$

15. *An insurance company has the following previous data on aggregate claims:*

Policyholder	*Year 1*	*Year 2*	*Year 3*	*Year 4*	*Mean*	*Variance*
1	*1,210*	*246*	*459*	*1,461*	*944.00*	*340158.00*
2	*0*	*0*	*0*	*0*	*0.00*	*0.00*
3	*0*	*2,185*	*0*	*0*	*548.25*	*1202312.25*
4	*809*	*0*	*0*	*1,725*	*633.50*	*674939.00*
5	*0*	*0*	*0*	*0*	*0.00*	*0.00*

Calculate the Bühlmann credibility premium for policyholder 3 in Year 5.

The expected process variance is $\frac{1}{5}(340158 + 0 + 1202312.25 + 674939 + 0) = 443421.85$. The population mean is $\frac{944+0+548.25+633.50+0}{5} = 405.15$.

total variance of estimated means is $\frac{(944-405.15)^2+(-405.15)^2+(548.25-405.15)^2+(633.50-405.15)^2+(-405.15)^2}{4} = 172318.425$. The variance of hypothetical means is therefore $172318.425 - \frac{443421.85}{4} = 61462.96$. The credibility of 4 years of experience is therefore $\frac{4}{4+\frac{443421.85}{61462.96}} = 0.3566825$. The premium for policyholder 3 is therefore $0.3566825 \times 548.25 + 0.6433175 \times 405.15 = \$456.19$.

16. *An insurance company collects the following claim frequency data for 7,000 customers insured for the past 3 years:*

No. of claims	*Frequency*
0	*1,491*
1	*2,461*
2	*1,810*
3	*831*
4	*302*
5	*72*
6	*30*
7	*2*
8	*1*
> 8	*0*

It assumes that the number of claims an individual makes in a year follows a Poisson distribution with parameter Λ, which may vary between individuals.

Find the credibility estimate for the expected number of claims per year for an individual who has made 4 claims in the past 3 years.

The total number of claims in the past 3 years was $1\times2461+2\times1810+3\times831+4\times302+5\times72+6\times30+7\times2+8\times1 = 9,345$. The total number of policyholders is $1491 + 2461 + 1810 + 831 + 302 + 72 + 30 + 2 + 1 = 7,000$. The average

number of claims per policyholder per year is therefore $\frac{9345}{21000} = 0.445$. This is also the expected process variance. The variance of estimated means is

$$\frac{1491 \times 0.445^2 + 2461\left(\frac{1}{3} - 0.445\right)^2 + 1810\left(\frac{2}{3} - 0.445\right)^2 + 826(1 - 0.445) + 307\left(\frac{4}{3} - 0.445\right)^2 + 72\left(\frac{5}{3} - 0.445\right)^2 + 30(2 - 0.445)}{6999}$$

The variance due to the Poisson sampling is $\frac{0.445}{3} = 0.148333$. Therefore, the variance of hypothetical means is $0.1553991 - 0.148333 = 0.0070658$. The credibility of 3 year's experience is $\frac{3}{3+\frac{0.445}{0.0070658}} = 0.04546872$. The expected number of claims is therefore $0.04546872 \times \frac{4}{3} + 0.95453128 \times 0.445 = 0.4853914$.

17. *Use the method of inversion to simulate two random samples from a Pareto distribution with $\alpha = 4$, $\theta = 6,200$.*

If U follows a uniform distribution, then the method of inversion gives a sample from a Pareto distribution by solving

$$\begin{aligned}
1 - \left(\frac{6200}{6200 + X}\right)^4 &= U \\
\left(\frac{6200}{6200 + X}\right)^4 &= 1 - U \\
\frac{6200}{6200 + X} &= (1 - U)^{\frac{1}{4}} \\
\frac{X}{6200} + 1 &= (1 - U)^{-\frac{1}{4}} \\
X &= 6200((1 - U)^{-\frac{1}{4}} - 1)
\end{aligned}$$

Substituting $U = 0.58665797$ and $U = 0.12487271$ gives $X = 1532.400$ and $X = 210.234$.

18. *An insurance company classifies individuals into three classes, each with a different claim severity distribution, as shown in the following table:*

Class	*Probability*	*Severity Distribution*	*Parameters*
1	*0.20*	*Pareto*	$\alpha = 4$, $\theta = 7,000$
2	*0.35*	*Weibull*	$\tau = 1.7$, $\theta = 800$
3	*0.45*	*Inverse Weibull*	$\tau = 2.8$, $\theta = 590$

Simulate 2 claim severities from 2 random individuals.

For the first severity, we first simulate to determine which component of the mixture applies. Our random value is 0.58665797, which is between 0.55 and 1, so the third component of the mixture applies, which is the Inverse Weibull distribution. We then use 0.12487271 to generate a random number by solving

$$
\begin{aligned}
e^{-\left(\frac{590}{X}\right)^{2.8}} &= 0.12487271 \\
\left(\frac{590}{X}\right)^{2.8} &= -\log(0.12487271) \\
\frac{590}{X} &= (-\log(0.12487271))^{\frac{1}{2.8}} \\
X &= \frac{590}{(-\log(0.12487271))^{\frac{1}{2.8}}} = 454.1754
\end{aligned}
$$

For the second random variable, we use 0.87530540 to simulate the component of the mixture. Again, it is between 0.55 and 1, so we simulate from the third component. We use 0.49197147 to simulate a random number by solving

$$
\begin{aligned}
e^{-\left(\frac{590}{X}\right)^{2.8}} &= 0.49197147 \\
\left(\frac{590}{X}\right)^{2.8} &= -\log(0.49197147) \\
\frac{590}{X} &= (-\log(0.49197147))^{\frac{1}{2.8}} \\
X &= \frac{590}{(-\log(0.49197147))^{\frac{1}{2.8}}} = 666.9902
\end{aligned}
$$

19. A pension plan has three types of exit with probabilities in the table below:

Exit Type	*Probability*
Retirement	*0.65*
Withdrawl	*0.25*
Death	*0.10*

Simulate the number of each type from a sample of 634 plan members. [You may use a normal approximation to the binomial distribution.]

The number who exit through retirement follows a binomial distribution with $n = 634$, and $p = 0.65$, which is approximately a normal distribution with mean $634 \times 0.65 = 412.1$, and variance $634 \times 0.65 \times 0.35 = 144.235$, which is standard deviation 12.00979. We use 0.58665797 to simulate from the normal distribution, we get $\Phi^{-1}(0.58665797) = 0.2189563$, so the simulated binomial random variable is $412.1 + 0.2189563 \times 12.00979 = 414.7296$, which after rounding becomes 415.

With 415 retirees, there are 219 members who withdraw or die. The conditional distribution of the number of withdrawls is therefore Binomial with $n = 219$ and $p = \frac{0.25}{0.35} = \frac{5}{7}$. The distribution is approximately normal with mean $219 \times \frac{5}{7} = 152.42856$ and variance $219 \times \frac{5}{7} \times \frac{2}{7} = 44.69388$, or standard deviation 6.685348. We have $\Phi^{-1}(0.12487271) = -1.150968$, so the simulated number of withdrawls is $152.42856 - 1.150968 \times 6.685348 = 148.7339$, which gives 149 withdrawls. The simulated number of deaths is therefore $219 - 149 = 70$.

20. Use a stochastic process method to simulate 2 samples from each of the following distributions:

(a) A Poisson distribution with $\lambda = 3$.

A Poisson distribution with $\lambda = 3$ is the number of exponential distributions with $\lambda = 3$ that can be added without exceeding 1. For a uniform random variable U, we can generate an exponential random variable T with parameter λ by solving:

$$\begin{aligned} 1 - e^{-\lambda T} &= U \\ e^{-\lambda T} &= 1 - U \\ T &= \frac{-\log(1-U)}{\lambda} \end{aligned}$$

Using $\lambda = 3$ and the given uniform random variables, we simulate

$$\begin{aligned} T_1 &= \frac{-\log(1-0.58665797)}{3} = 0.29449329 \\ T_2 &= \frac{-\log(1-0.12487271)}{3} = 0.04446198 \\ T_3 &= \frac{-\log(1-0.87530540)}{3} = 0.69396258 \end{aligned}$$

This gives $T_1 + T_2 + T_3 = 1.0329178 > 1$, so the first simulated number is 2.

Next we simulate

$$\begin{aligned} T_1 &= \frac{-\log(1-0.49197147)}{3} = 0.22573922 \\ T_2 &= \frac{-\log(1-0.55262301)}{3} = 0.26811789 \\ T_3 &= \frac{-\log(1-0.14644543)}{3} = 0.05278193 \\ T_4 &= \frac{-\log(1-0.89151074)}{3} = 0.74036803 \end{aligned}$$

This gives $T_1 + T_2 + T_3 + T_4 = 1.287007 > 1$, so the simulated value is 3.

(b) A negative binomial distribution with $r = 7$ and $\beta = 0.52$.

For the negative binomial, we simulate $\lambda_k = (7 + k)\log(1.52) = 2.930972 + 0.4187103k$.

We then simulate

$$T_1 = \frac{-\log(1-0.58665797)}{2.930972} = 0.30142893$$
$$T_2 = \frac{-\log(1-0.12487271)}{3.349683} = 0.03982047$$
$$T_3 = \frac{-\log(1-0.87530540)}{3.768393} = 0.55246035$$
$$T_4 = \frac{-\log(1-0.49197147)}{4.187103} = 0.16173894$$

This gives $T_1 + T_2 + T_3 + T_4 = 1.0554487 > 1$, so the simulated value is 3.

For the second simulated value, we simulate:

$$T_1 = \frac{-\log(1-0.5526230)}{2.930972} = 0.27443236$$
$$T_2 = \frac{-\log(1-0.1464454)}{3.349683} = 0.04727188$$
$$T_3 = \frac{-\log(1-0.8915107)}{3.768393} = 0.58940352$$
$$T_4 = \frac{-\log(1-0.4655928)}{4.187103} = 0.14964931$$

This gives $T_1 + T_2 + T_3 + T_4 = 1.060757 > 1$, so the simulated value is 3.

21. Simulate 2 samples from a normal distribution with $\mu = 3$ and $\sigma = 7$ using

(a) A Box-Muller transformation.

Using the Box-Muller transformation, we simulate

$$Z_1 = \sqrt{-2\log(0.58665797)}\cos(0.12487271 \times 2\pi) = 0.7308669$$
$$Z_2 = \sqrt{-2\log(0.58665797)}\sin(0.12487271 \times 2\pi) = 0.7296987$$
$$X_1 = 7Z_1 + 3 = 8.116068$$
$$X_2 = 7Z_2 + 3 = 8.107891$$

(b) The polar method.

Using the polar method, we get $X_1 = 2 \times 0.58665797 - 1 = 0.17331594$ and $X_2 = 2 \times 0.12487271 - 1 = -0.75025458$. Now we get $W = {X_1}^2 + {X_2}^2 = 0.17331594^2 + (-0.75025458)^2 = 0.5929203$. Since $W < 1$, we do not need to simulate new values. We now let $Y = \sqrt{\frac{-2\log(W)}{W}} = 1.327826$. The simulated values are then $YX_1 = 1.327826 \times 0.17331594 = 0.2301334$ and $YX_2 = 1.327826 \times -0.75025458 = -0.9962073$.

$$7YX_1 + 3 = 4.610934$$
$$7YX_2 + 3 = -9.973451$$

22. *An insurance company is simulating its aggregate losses. It is attempting to estimate the probability that its aggregate losses exceed \$1,000,000.*

(a) How many aggregate losses does it need to simulate to ensure that there is a 99% probability that the estimated probability of exceeding \$1,000,000 is within 0.001 of the true probability, regardless of the true probability?

If the true probability is p, then the estimated probability from n simulations is a scaled binomial distribution with parameters n and p. This can be approximated by a normal distribution with mean p and variance $\frac{p(1-p)}{n}$. The probability that the estimated probability is within 0.001 of the true probability is $2\Phi\left(\frac{0.001}{\sqrt{\frac{p(1-p)}{n}}}\right) - 1$. Setting this to 0.99 gives

$$\begin{aligned}
2\Phi\left(\frac{0.001}{\sqrt{\frac{p(1-p)}{n}}}\right) - 1 &= 0.99 \\
\Phi\left(\frac{0.001}{\sqrt{\frac{p(1-p)}{n}}}\right) &= 0.995 \\
\frac{0.001}{\sqrt{\frac{p(1-p)}{n}}} &= 2.575829 \\
\sqrt{\frac{p(1-p)}{n}} &= \frac{0.001}{2.575829} \\
n &= 2575.829^2 p(1-p)
\end{aligned}$$

We see that this is maximised when $p = 0.5$, which gives $n = 1,658,724$.

(b) Suppose the true probability of aggregate losses exceeding \$1,000,000 is 0.05. How many simulations does the company need to perform in order for the relative error in this estimated probability to be less than 1% with probability 0.95?

If the true probability is 0.05, then a relative error less than 1% means that the estimated probability is between 0.0495 and 0.0505. We want the probability of this to be 0.95. Using the same normal approximation, with $p = 0.05$ gives:

$$2\Phi\left(\frac{0.0005}{\sqrt{\frac{0.05\times 0.95}{n}}}\right) - 1 = 0.95$$

$$\Phi\left(\frac{0.0005}{\sqrt{\frac{0.05\times 0.95}{n}}}\right) = 0.975$$

$$\frac{0.0005}{\sqrt{\frac{0.05\times 0.95}{n}}} = 1.96$$

$$n = 3920^2 \times 0.0475 = 729904$$

23. A reinsurance company is using a simulation to calculate the premium for a stop-loss insurance contract. It simulates 100,000 outcomes, and finds that the mean payment is $492,384, and the standard deviation of the payments is $2,643,000. It wants to calculate the net premium with a 99% chance that the relative error in its net premium is less than 1%. Assuming the mean and standard deviation are similar to the results it already has, how many more simulations does it need to perform to achieve this accuracy?

By the central limit theorem, the mean of the simulation can be approximated by a normal distribution with mean the true mean, which is approximately 492,384, and variance $\frac{2643000^2}{n}$. If these are the true mean and variance, then to have a 1% relative error with 99% probability, it needs to simulate a total of n simulations so that

$$2\Phi\left(\frac{0.01 \times 492384}{\left(\frac{2643000}{\sqrt{n}}\right)}\right) - 1 = 0.99$$

$$\Phi\left(\frac{4923.84\sqrt{n}}{2643000}\right) = 0.995$$

$$0.001862974\sqrt{n} = 2.575829$$

$$n = 1,911,704$$

So it needs to perform another 1,811,704 simulations.

24. An insurance company is estimating its aggregate losses. It simulates 1000 claim frequencies, and finds a total of 749 claims. It therefore simulates 749 claim severities, and simulates the aggregate losses by adding the claim severities in groups corresponding to the simulated claim frequencies. The insurance company has a second line of insurance which also has the same severity distribution, but a different frequency distribution. It simulates 1000 new frequencies and gets a total of 749 claims again. It uses the same simulated claim severities to model aggregate losses for the second line of insurance. Based on these simulated values, it calculates a 95% confidence interval for the aggregate losses. Which of the following statements best describes this procedure? Explain your answer.

(i) The procedure is sound and should produce an accurate confidence interval.

(ii) The procedure is unsound and will produce a narrower confidence interval than it should (so the confidence interval will contain the true value less than 95% of the time).

(iii) The procedure is unsound and will produce a wider confidence interval than it should (so the confidence interval will contain the true value more than 95% of the time).

(iv) The procedure is unsound, and the confidence interval will be wider than it should in some cases and narrower than it should in others.

Using the same simulated values for the claim severities will mean that the aggregate losses for the two lines are correlated. The assumption that the variance is decreased by taking the average of two lines of insurance is not correct because they are using the same values for the simulation, so they are correlated. This will make the confidence interval narrower than it should be. On the other hand, if the variance of the aggregate losses is calculated based on the variance of the sum of the aggregate losses from the two lines of insurance, then using the same aggregate losses may create correlation between these values, so that it has more estimated variance than it should, and therefore, will give a wider confidence interval than it should. The extent to which this happens will depend how the simulated frequencies correspond. Therefore, whether the procedure will produce a wider or narrower confidence interval than it should will vary between simulations. Therefore (iv) best describes the procedure.