ACSC/STAT 4703, Actuarial Models I
Further Probability with Applications to Actuarial Science
Fall 2015

Toby Kenney

In Class Examples

Advantages of Modelling Number of Losses and Severities Separately

- Dealing with changes to exposure (e.g. number of policies)
- Dealing with inflation
- Dealing with changes to individual policies
- Understanding the impact of changing deductibles on claim frequencies.
- Combining data with a range of different deductibles and limits can give a better picture of the loss distribution.
- Consistency between models of non-covered losses to insureds, claims to insurers, and claims to reinsurers.
- The effect of the shapes of separate distributions of number and severity give an indicator of how each influences the overall aggregate loss.

Model Choices

Practical Considerations

- Scale parameters for severity allow for change of currency or inflation.
- For frequency, models with pgf $P(z;\alpha) = Q(z)^{\alpha}$ can deal with changes to number of policies sold, or time period.
- Modification at zero prevents infinite divisibility. However, modification at zero may still be appropriate.

Question 1

Which discrete distributions satisfy

$$P(z; \alpha) = Q(z)^{\alpha}$$

for some parameter α?

Question 2

Calculate the first three moments of a compound model.

Question 3

When an individual is admitted to hospital, the distribution of charges incurred are as described in the following table:

charge	mean	standard deviation
Room	1000	500
other	500	300

The covariance between room charges and other charges is 100,000. An insurer issues a policy which reimburses 100% for room charges and 80% for other charges. The number of hospital admissions has a Poisson distribution with parameter 4. Determine the mean and standard deviation for the insurer's payout on the policy.

Question 4

An individual loss distribution is normal with mean 100 and standard deviation 35. The total number of losses N has the following distribution:

n	$P(N = n)$
0	0.4
1	0.3
2	0.2
3	0.1

What is the probability that the aggregate losses exceed 130?

Question 5

Aggregate payments have a compound distribution. The frequency distribution is negative binomial with $r = 16$, $\beta = 6$. The severity distribution is uniform on the interval $(0, 8)$. Using a normal approximation, determine the premium such that there is a 5% probabilty that aggregate payments exceed the premium.

Question 6

For a group health contract, aggregate claims are assumed to have an exponential distribution with mean θ estimated by the group underwriter. Aggregate stop-loss insurance for total claims in excess of 125% of the expected claims, is provided for a premium of twice the expected stop-loss claims. It is discovered that the expected total claims value used was 10% too low. What is the loading percentage on the stop-loss policy under the true distribution?

Question 7

Calculate the probability density function of the aggregate loss distribution if claim frequency follows a negative binomial distribution and severity follows an exponential distribution.

Question 8

An insurance company models the number of claims it receives as a negative binomial distribution with parameters $r = 15$ and $\beta = 2.4$. The severity of each claim follows an exponential distribution with mean \$3,000. What is the net-premium for stop-loss insurance with a deductible of \$500,000?

Question 9

An insurance company offers group life insurance policies to three different companies. For the first company, the number of claims is a Poisson distribution with parameter $\lambda = 0.4$, and claim severity a gamma distribution with $\theta = 30,000$ and $\alpha = 3$. For the second company, the number of claims is a Poisson distribution with parameter $\lambda = 3.6$ and the severity follows a gamma distribution with $\theta = 200,000$ and $\alpha = 1.4$. For the third company, the number of claims follows a Poisson distribution with $\lambda = 85$ and claim severity follows a gamma distribution with $\theta = 45,000$ and $\alpha = 2.2$. What is the probability that the aggregate claims from all these policies exceed 10,000,000?

Question 10

Suppose that the total number of claims follows a negative binomial distribution with $r = 2$ and $\beta = 3$. Suppose that the severity of each claim (in thousands of dollars) follows a zero-truncated ETNB distribution with $r = -0.6$ and $\beta = 7$. What is the probability that the aggregate loss is at most 3?

The Recursive Method

Theorem

Suppose the severity distribution is a discrete distribution with probability function $f_X(x)$ *for* $x = 0, 1, \ldots, m$ *(m could be infinite) and the frequency distribution is a member of the* $(a, b, 1)$ *class with probabilities* $p_k, k = 0, 1, 2, \ldots$ *satisfying* $p_k = \left(a + \frac{b}{k}\right) p_{k-1}$ *for all* $k \geqslant 2$*.*
Then the aggregate loss distribution is given by

$$f_S(x) = \frac{(p_1 - (a+b)p_0)f_X(x) + \sum_{y=1}^{x \wedge m} \left(a + \frac{by}{x}\right) f_X(y) f_S(x-y)}{1 - af_X(0)}$$

Question 11

Let the number of claims follow a Poisson distibution with $\lambda = 2.4$ and the severity of each claim follow a negative binomial distribution with $r = 10$ and $\beta = 2.3$. What is the probability that the aggregate loss is at most 3?

Question 12

An insurance company offers car insurance. The number of losses a driver experiences in a year follows a negative binomial random variable with $r = 0.2$ and $\beta = 0.6$. The size of each loss (in hundreds of dollars) is modelled as following a zero-truncated ETNB distribution with $r = -0.6$ and $\beta = 3$. The policy has a deductible of \$1,000 per loss. What is the probability that the company has to pay out at least \$400 in a single year to a driver under such a policy?

Question 13

The number of claims an insurance company receives is modelled as a compound Poisson distribution with parameter $\lambda = 6$ for the primary distribution and $\lambda = 0.1$ for the secondary distribution. Claim severity (in thousands of dollars) is modelled as following a zero-truncated logarithmic distribution with parameter $\beta = 4$. What is the probability that the total amount claimed is more than $3,000.

Question 14

The number of claims an insurance company receives is modelled as a Poisson distribution with parameter $\lambda = 96$. The size of each claim is modelled as a zero-truncated negative binomial distribution with $r = 4$ and $\beta = 2.2$. Calculate the approximated distribution of the aggregate claims:

(a) By starting the recursion at a value of k six standard deviations below the mean.

(b) By solving for a rescaled Poisson distribution with $\lambda = 12$ and convolving the solution up to 96.

Answer to Question 14

R-Code for (a)

```
ans<-1
ans<-as.vector(ans)
for(n in 2:2000){
  temp<-0
  for(i in 1:(n-1)){%
    temp<-temp+16*i*(i+1)*(i+2)*(i+3)/(n+240)*0.6875^i*
        0.3125^4*ans[n-i]/(1-0.3125^4)
  }
  ans<-c(ans,temp)
}
```

Answer to Question 14

R-Code for (b)

```
ConvolveSelf<-function(n){
  convolution<-vector("numeric",2*length(n))
  for(i in 1:(length(n))){
    convolution[i]<-sum(n[1:i]*n[i:1])
  }
  for(i in 1:(length(n))){
    convolution[2*length(n)+1-i]<-sum(n[length(n)+1-(1:i)
        ]*n[length(n)+1-(i:1)])
  }
  return(convolution)
}

d24<-ConvolveSelf(ans2)
d48<-ConvolveSelf(d24)
d96<-ConvolveSelf(d48)
plot(dist1,d96[241:2240])
```

Question 15

Let X follow an exponential distribution with mean θ. Approximate this with an arithmetic distribution ($h = 1$) using:
(a) The method of rounding.
(b) The method of local moment matching, matching 2 moments on each interval.

Question 16

The loss on a given policy is modelled as following an exponential distribution with mean 2,000. The number of losses follows a negative binomial distribution with parameters $r = 4$ and $\beta = 2.1$.
(a) Calculate the distribution of the aggregate loss.
(b) What effect would a deductible of \$500 have on this distribution?

Question 17

The loss on a given policy is modelled as following a gamma distribution with $\alpha = 3.4$ and $\theta = 2000$. The number of losses an insurance company insures follows a Poisson distribution with $\lambda = 20$. The company has taken out stop-loss insurance with a deductible of \$1,000,000. This insurance is priced at the expected payment on the policy plus one standard deviation.
(a) How much does the company pay for this reinsurance?
(b) How much should it pay if it introduces a \$1,000 deductible on these policies?

Question 18

In a group life insurance policy, a life insurance company insures 10,000 individuals at a given company. It classifies these workers in the following classes:

Type of worker	number	average annual probability of dying	average death benefit
Manual Laborer	4,622	0.01	\$100,000
Administrator	3,540	0.002	\$90,000
Manager	802	0.01	\$200,000
Senior Manager	36	0.02	\$1,000,000

What is the probability that the aggregate benefit paid out in a year exceeds \$10,000,000?

Question 19

Using the same data as in Question 18, estimate the probability by modelling the distribution of the aggregate risk as:
(a) a normal distribution
(b) a gamma distribution
(c) a log-normal distribution

Question 20

Using the same data as in Question 18, estimate the probability using a compound Poisson approximation, setting the Poisson mean to:
(a) equal the Bernoulli probability
(b) match the probability of no loss

Question 21

An insurance company has the following portfolio of car insurance policies:

Type of driver	Number	Probability claim	mean of claim	standard deviation
Safe drivers	800	0.02	\$3,000	\$1,500
Average drivers	2100	0.05	\$4,000	\$1,600
Dangerous drivers	500	0.12	\$5,000	\$1,500

(a) Using a gamma approximation for the aggregate losses on this portfolio, calculate the cost of reinsuring losses above \$800,000, if the loading on the reinsurance premium is one standard deviation above the expected claim payment on the reinsurance policy.
(b) How much does the premium change if we use a normal approximation?

Question 22

An insurance company assumes that for smokers, the claim probability is 0.02, while for non-smokers, it is 0.01. A group of mutually independant lives has coverage of 1000 per life. The company assumes that 20% of lives are smokers. Based on this assumption, the premium is set equal to 110% of expected claims. If 30% of the lives are smokers, the probability that claims will exceed the premium is less than 0.2. Using a normal approximation, determine the minimum number of lives in the group.

Question 23

Calculate the empirical distribution and cumulative hazard rate function for the following data set:

4 3 6 0 3 7 0 0 2 1 1 3 6 3 4

11.2 The Empirical Distribution for Complete Individual Data

Question 24

For the data set from the previous question,

4 3 6 0 3 7 0 0 2 1 1 3 6 3 4

compute a Nelson-Åalen estimate for the probability that a random sample is larger than 5.

Question 25

An insurance company collects the following data on life insurance policies:

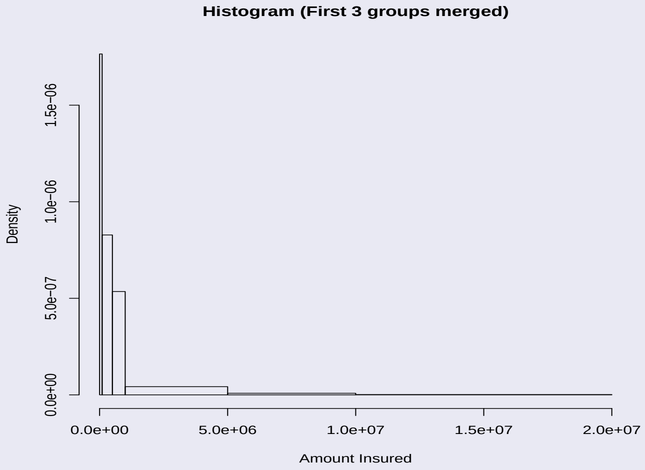
Amount Insured	Number of Policies
Less than \$5,000	30
\$5,000–\$20,000	52
\$20,000–\$100,000	112
\$100,000–\$500,000	364
\$500,000–\$1,000,000	294
\$1,000,000–\$5,000,000	186
\$5,000,000–\$10,000,000	45
More than \$10,000,000	16

The government is proposing a tax on insurance policies for amounts larger than \$300,000. Using the ogive to estimate the empirical distribution function, what is the probability that a random policy is affected by this tax?

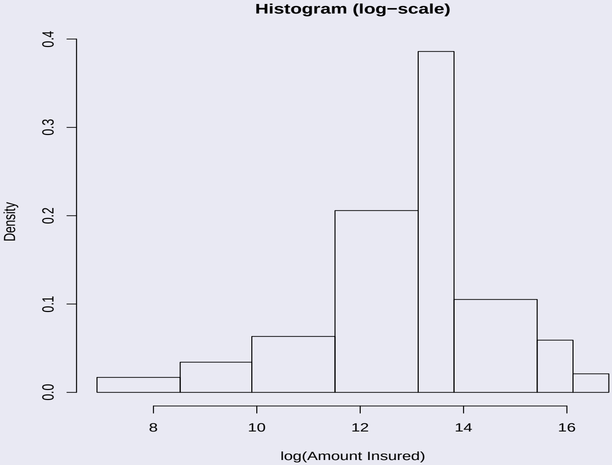
Question 26

Draw the histogram for the data from Question 25

Answer to Question 26



Answer to Question 26



Question 27

A sample of size 2,000 contains 1,700 observations that are at most 6,000; 30 that are between 6,000 and 7,000; and 270 that are more than 7,000. The total of the 30 observations between 6,000 and 7,000 is 200,000. The value of $\mathbb{E}(X \wedge 6000)$ under the empirical distribution obtained from this data is 1,810. Calculate the value of $\mathbb{E}(X \wedge 7000)$ under the empirical distribution obtained from this data.

Question 28

A random sample of unknown size includes 36 observations between 0 and 50, x observations between 50 and 150, y observations between 150 and 250, 84 observations between 250 and 500, 80 observations between 500 and 1,000, and no observations above 1,000.
The ogive includes the values $F_n(90) = 0.21$ and $F_n(210) = 0.51$.
Calculate x and y.

12.1 Point Estimation— Truncation and Censoring

Definition

Truncated from below	Observations $\leqslant d$ are not recorded
Truncated from above	Observations $\geqslant u$ are not recorded
Censored from below	Observations $\leqslant d$ are recorded only as $\leqslant d$
Censored from above	Observations $\geqslant u$ are recorded only as $\geqslant u$

We will by default assume truncation is on the left (deductible) and censoring is on the right (policy limit).

Example Data Set

i	d_i	x_i	u_i	i	d_i	x_i	u_i	i	d_i	x_i	u_i
1	0	0.4	-	8	0	-	1.8	15	1.2	-	1.4
2	0	1.6	-	9	0	1.4	-	16	0.5	-	1.3
3	0	-	2.4	10	0	-	1.2	17	0.5	2.2	-
4	0	0.7	-	11	0	1.3	-	18	0.9	-	2.3
5	0	-	0.4	12	0	-	1.1	19	0.8	1.2	-
6	0	1.9	-	13	0.4	1.4	-	20	0.6	-	1.5
7	0	1.1	-	14	0.7	1.7	-	21	1.1	1.8	-

Notes

- d_i is the left truncation point (deductible). 0 indicates no truncation.
- x_i indicate complete data. That is the exact value of x_i is known.
- u_i indicate censored data. All that is known is that $x_i \geqslant u_i$.
- There should be an entry in exactly one of the third and fourth columns.

Kaplan-Meier Product-Limit Estimator

Notation

- y_j Unique uncensored values sorted in increasing order $y_1 < \ldots < y_k$
- s_j Number of times y_j occurs in the sample
- r_j Size of risk set at y_j. That is the number of samples i such that $d_i < y_j < x_i$ or $d_i < y_j < u_i$

Formulae

$$r_j = |\{i|x_i \geqslant r_j\}| + |\{i|u_i \geqslant r_j\}| - |\{i|d_i \geqslant r_j\}|$$
$$r_j = |\{i|d_i < r_j\}| - |\{i|u_i < r_j\}| - |\{i|x_i < r_j\}|$$

Kaplan-Meier Product Limit-Estimator

For $y_{j-1} < t \leqslant y_j$:

$$S(t) = \prod_{i=1}^{j-1} \left(1 - \frac{s_i}{r_i}\right)$$

Summary of Dataset from Previous Slide

i	y_i	s_i	r_i	i	y_i	s_i	r_i	i	y_i	s_i	r_i
1	0.4	1	12	5	1.3	1	13	9	1.8	1	6
2	0.7	1	14	6	1.4	2	11	10	1.9	1	4
3	1.1	1	16	7	1.6	1	8	11	2.2	1	3
4	1.2	1	15	8	1.7	1	7				

Question 29

Using the summary of the dataset above, and the Kaplan-Meier product-limit estimator, estimate the probability that a randomly chosen observation is more than 1.6.

Question 30

Using the data in the table:

i	d_i	x_i	u_i	i	d_i	x_i	u_i	i	d_i	x_i	u_i
1	0	1.3	-	8	0	0.8	-	15	0.2	0.6	-
2	0	0.1	-	9	0	-	0.7	16	0.6	1.4	-
3	0	-	0.4	10	0.2	-	0.8	17	0.2	0.3	-
4	0	0.2	-	11	0.3	1.4	-	18	0.1	-	1.0
5	0	-	0.3	12	0.3	-	0.8	19	0.1	0.3	-
6	0	0.8	-	13	0.2	-	0.5	20	0.3	1.3	-
7	0	-	0.1	14	0.4	-	0.5	21	0.3	1.4	-

Using the Kaplan-Meier product-limit estimator, estimate the median of the distribution.

Answer to Question 30

i	y_i	s_i	r_i
1	0.1	1	12+9-12= 9
2	0.2	1	11+8-10= 9
3	0.3	2	10+8- 6=12
4	0.6	1	8+4- 1=11
5	0.8	2	7+3- 0=10
6	1.3	2	5+0- 0= 5
7	1.4	3	3+0- 0= 3

Question 31

For the same data as in Question 29, summarised in the following table:

i	y_i	s_i	r_i	i	y_i	s_i	r_i	i	y_i	s_i	r_i
1	0.4	1	12	5	1.3	1	13	9	1.8	1	6
2	0.7	1	14	6	1.4	2	11	10	1.9	1	4
3	1.1	1	16	7	1.6	1	8	11	2.2	1	3
4	1.2	1	15	8	1.7	1	7				

using a Nelson-Åalen estimator, estimate the probability that a random observation is larger than 1.6.

Question 32

From the data in Question 30 summarised below:

i	y_i	s_i	r_i
1	0.1	1	12+9-12= 9
2	0.2	1	11+8-10= 9
3	0.3	2	10+8- 6=12
4	0.6	1	8+4- 1=11
5	0.8	2	7+3- 0=10
6	1.3	2	5+0- 0= 5
7	1.4	3	3+0- 0= 3

Estimate the probability that a random observation that is known to be more than 0.5 is at most 1. Use a Kaplan-Meier estimator.

Question 33

Calculate the variance of the empirical survival function for grouped data using the ogive.

Question 34

An insurance company receives 4,356 claims, of which 2,910 are less than \$10,000, and 763 are between \$10,000 and \$100,000. Calculate a 95% confidence interval for the probability that a random claim is larger than \$50,000.

Question 35

Show that under the assumption that the sizes of the risk set and the possible dying times are fixed, the Kaplan-Meier product-limit estimate is unbiassed and calculate its variance.

Greenwood's Approximation

Approximation

If $a_1, \ldots, a_n$ are all small, then

$$(1 + a_1) \cdots (1 + a_n) \approx 1 + a_1 + a_2 + \cdots + a_n$$

Formula

$$\mathrm{Var}(S_n(y_j)) \approx \left(\frac{S(y_j)}{S(y_0)} \right)^2 \sum_{i=1}^{j} \frac{S(y_{i-1}) - S(y_i)}{r_i S(y_i)}$$

Since $\frac{r_i - s_i}{r_i}$ is an estimate of $\frac{S(y_i)}{S(y_{i-1})}$, we can estimate this by

$$\mathrm{Var}(S_n(y_j)) \approx \hat{S}(y_j)^2 \sum_{i=1}^{j} \frac{s_i}{r_i(r_i - s_i)}$$

Question 36

Recall from Question 30 (data summarised below) that using a Kaplan Meier estimator, we have

$$S_n(1) = \frac{8}{9} \times \frac{8}{9} \times \frac{10}{12} \times \frac{10}{11} \times \frac{8}{10} = \frac{1280}{2673}$$

Use Greenwood's formula to find a 95% confidence interval for $S_n(1)$

i	y_i	s_i	r_i
1	0.1	1	12+9-12= 9
2	0.2	1	11+8-10= 9
3	0.3	2	10+8- 6=12
4	0.6	1	8+4- 1=11
5	0.8	2	7+3- 0=10
6	1.3	2	5+0- 0= 5
7	1.4	3	3+0- 0= 3

Log-transformed Confidence Interval

Problem

The usual method for constructing a confidence interval for $S(x)$ may lead to impossible values (negative or more than 1).

Solution

Instead find a confidence interval for

$$\log(-\log(S(x)))$$

which has no impossible values.

Log-transformed Confidence Interval (Continued)

Method

By the delta method, if $S_n(x)$ is approximately normal with mean μ and small variance σ^2, then for any smooth function $g(x)$, we have that $g(S_n(x))$ is approximately normal with mean $g(\mu)$ and variance $g'(\mu)^2\sigma^2$.
In particular, when $g(x) = \log(-\log(S(x)))$, we have $g'(x) = \frac{1}{S(x)\log(S(x))}$.

Definition

The log-transformed confidence interval for $S(X)$ is given by $[S_n(x)^{\frac{1}{U}}, S_n(X)^U]$, where $U = e^{\Phi^{-1}\left(\frac{\alpha}{2}\right)\frac{\sigma}{S_n(x)\log(S_n(x))}}$.

Question 37

Recall from Question 36 that for the following data set

i	y_i	s_i	r_i
1	0.1	1	12+9-12= 9
2	0.2	1	11+8-10= 9
3	0.3	2	10+8- 6=12
4	0.6	1	8+4- 1=11
5	0.8	2	7+3- 0=10
6	1.3	2	5+0- 0= 5
7	1.4	3	3+0- 0= 3

The Kaplan-Meier estimator is $S_n(1) = \frac{1280}{2673} = 0.4788627$ and Greenwood's formula gives the variance as 0.0180089. Find a 95% log-transformed confidence interval for $S(1)$.

Question 38

An insurance company observes the following claims

Claim size	Frequency	r_i
1	226	1641
2	387	1415
3	290	1028
4	215	738
5	176	523
7	144	347
9	97	203
> 9	106	

Use a Nelson Åalen estimator to obtain a 95% log-transformed confidence interval for the probability that a random claim is more than 5.

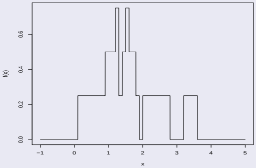
Question 39

An insurance company observes the following claims:

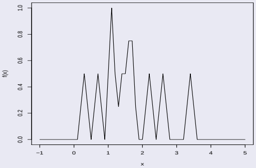
0.3	0.7	1.1	1.1	1.4	1.6	1.7	2.2	2.6	3.4	5.1

Estimate the probability density of the distribution using:
(a) A uniform kernel with bandwidth 0.2
(b) A uniform kernel with bandwidth 1.3
(c) A triangular kernel with bandwidth 0.2
(d) A triangular kernel with bandwidth 1.3
(e) A gamma kernel with $\alpha = 4.6$
(f) A gamma kernel with $\alpha = 1.6$

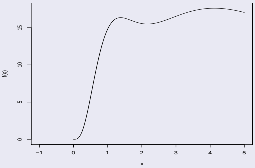
Answer to Question 39



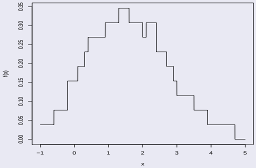
Uniform, bandwidth 0.2



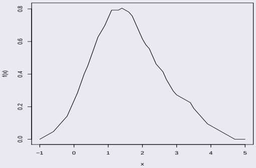
Triangle, bandwidth 0.2



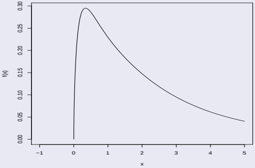
Gamma, $\alpha = 4.6$



Uniform, bandwidth 1.3



Triangle, bandwidth 1.3



Gamma, $\alpha = 1.6$

Question 40

Simulate *N* data points from a Pareto distribution with $\alpha = 3$ and $\theta = 4$. Use a kernel density estimator for each of the six kernels used in Question 39, and compare the kernel density estimate with the true distribution for varying sample sizes.

R-code for Question 40

```
kerneldensity<-function(data,kernel){
    f<-rep(0,6000)
    for(i in (1:6000)){
        f[i]<-mean(kernel(i/1000,data))
    }
    plottrue((1:6000)/1000)
    points((1:6000)/1000,f,type='l',col="red")
}

runSimulation<-function(N,kernel){
    data<-runif(N)
    alpha<-3
    theta<-4
    data<-(1/data^(1/alpha)-1)*theta
    kerneldensity(data,kernel)
}
```

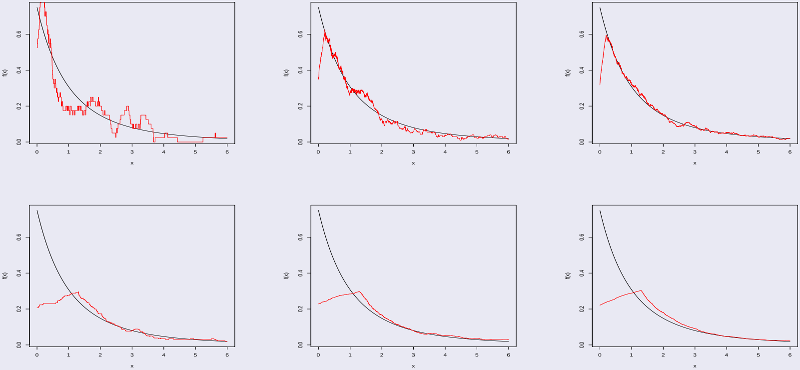
Answer to Question 40

Uniform Kernel

$N = 100$ $N = 500$ $N = 2000$

0.2

1.3



12.3 Kernel Density Models

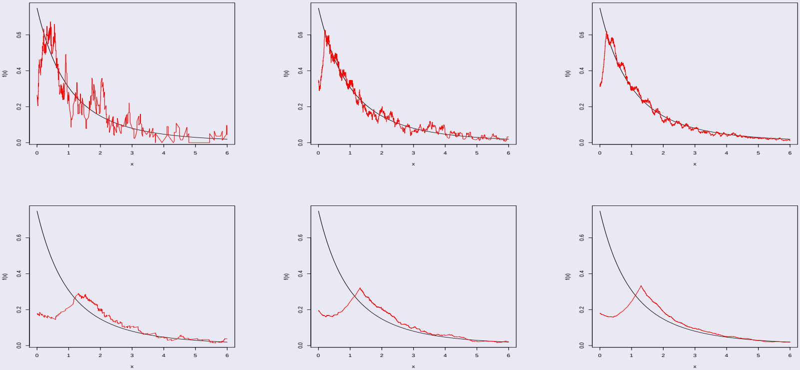
Answer to Question 40

Triangular Kernel

$N = 100$ $\quad$ $N = 500$ $\quad$ $N = 2000$

0.2

1.3



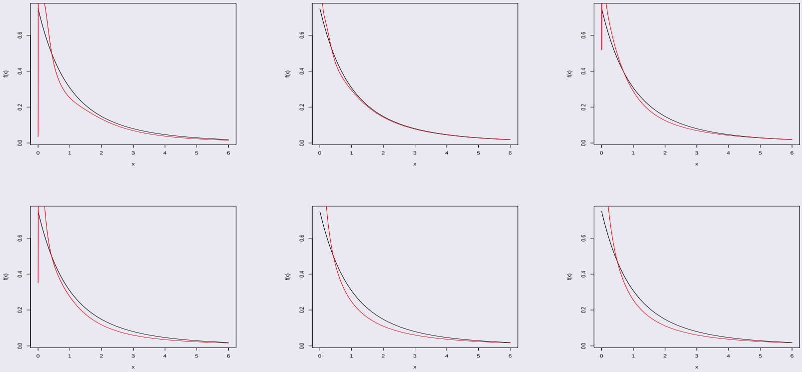
Answer to Question 40

Gamma Kernel

$N = 100$ $N = 500$ $N = 2000$

4.6

1.6



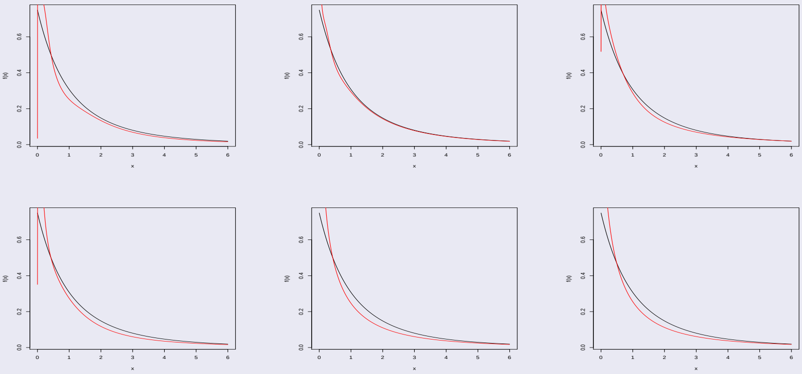
Answer to Question 40

$N = 100000$

Uniform Triangle Gamma

Low

High



Aim

Construct a lifetable from data in a mortality study. For each individual this data includes:

- Age at entry. (This might either be when the policy was purchased or when the study started if the policy was purchased before this time.)
- Age at exit.
- Reason for exit (death or other). Other exits might be surrender or termination of policy or end of study period.

Two Methods

Exact Exposure

- Exposure e_i of each observation during an age range is the proportion of that age range for which the individual was in the study.
- $\frac{d_i}{e_i}$ (deaths divided by exposures) is the estimator for hazard rate.
- The probability of dying within the age range is $1 - e^{-\frac{d_i}{e_i}t}$ where t is the length of the age range.

Actuarial Exposure

- Exposure e_i is the proportion of the age range for which the individual was either in the study or dead.
- That is, individuals who die are assumed to remain in the study until the end of the age range.
- $\frac{d_i}{e_i}t$ is the estimator for the probability of dying within the age range.

Question 41

An insurance company records the following data in a mortality study:

entry	death	exit	entry	death	exit	entry	death	exit
61.4	-	64.4	61.9	-	64.9	62.1	-	63.5
62.4	-	63.7	60.6	-	63.4	62.6	63.1	-
62.7	-	64.4	61.3	-	63.8	63.1	65.3	-
61.0	-	63.2	63.8	-	64.8	63.4	65.6	-
63.2	-	66.2	62.2	-	64.4	61.8	63.2	-
62.7	-	65.0	61.8	-	63.4	62.2	63.4	-
63.6	-	66.6	62.6	-	65.6			

Estimate the probability of an individual currently aged exactly 63 dying within the next year using:

(a) the exact exposure method.

(b) the actuarial exposure method.

12.4 Approximations for Large Data Sets

Insuring Ages

- Premiums based on whole ages only.
- q_{36} — the probability of an individual aged 36 dying within a year — is not for an individual aged exactly 36, but rather for an average individual aged 36.
- Now an individuals age is changed slightly so that their birthday is adjusted to match the date on which they purchased the policy.

Anniversary-based Mortality Studies

- Policyholders enter the study on the first policy anniversary following the start of the study.
- Policyholders leave the study on the last policy anniversary before the scheduled end of the study or their surrender.

Question 42

Recall the following data from Question 41 (with an additional column showing the age at which the individuals purchased their policy):

purchase	entry	death	exit	purchase	entry	death	exit
58.2	61.4	-	64.4	63.8	63.8	-	64.8
53.7	62.4	-	63.7	56.4	62.2	-	64.4
59.3	62.7	-	64.4	60.4	61.8	-	63.4
48.9	61.0	-	63.2	56.0	62.6	-	65.6
59.4	63.2	-	66.2	61.8	61.8	63.2	-
62.7	62.7	-	65.0	62.2	62.2	63.4	-
61.0	63.6	-	66.6	61.7	63.4	65.6	-
55.2	61.9	-	64.9	55.0	62.1	-	63.5
38.4	60.6	-	63.4	52.4	62.6	63.1	-
49.9	61.3	-	63.8	60.3	63.1	65.3	-

(a) Calculate the estimate for q_{63} using insuring ages.

(b) Now recalculate q_{63} on an anniversary-to-anniversary basis.

Question 43

Recall the data from Question 41:

entry	death	exit	entry	death	exit	entry	death	exit
61.4	-	64.4	61.9	-	64.9	62.1	-	63.5
62.4	-	63.7	60.6	-	63.4	62.6	63.1	-
62.7	-	64.4	61.3	-	63.8	63.1	65.3	-
61.0	-	63.2	63.8	-	64.8	63.4	65.6	-
63.2	-	66.2	62.2	-	64.4	61.8	63.2	-
62.7	-	65.0	61.8	-	63.4	62.2	63.4	-
63.6	-	66.6	62.6	-	65.6			

Rewrite the information from this table showing only the events by age interval.

Answer to Question 43

Age	Number at start	enter	die	leave	Number at next age
60	0	1	0	0	1
61	2	5	0	0	7
62	7	8	0	0	15
63	15	5	3	6	11
64	11	0	0	5	6
65	5	0	2	1	2
66	2	0	0	2	0

Question 44

Using the above table estimate q_{63} (the probability that an individual aged exactly 63 dies within one year). Assuming events are uniformly distributed over the year and use:

(a) exact exposure.

(b) actuarial exposure.

Question 45

Using the table from Question 44:

Age	Number at start	enter	die	leave	Number at next age
60	0	1	0	0	1
61	2	5	0	0	7
62	7	8	0	0	15
63	15	5	3	6	11
64	11	0	0	5	6
65	5	0	2	1	2
66	2	0	0	2	0

Estimate the probability that an individual aged exactly 63 withdraws from their policy within the next year conditional on surviving to age 64.

15 Bayesian Estimation

Prior Distribution

Before the experiment, we have beliefs about how plausible various values are for the parameter θ being estimated. These beliefs form a probability distribution called the prior distribution and denoted $\pi(\theta)$.

Theorem (Bayes Theorem)

For an event A and mutually exclusive events B_i with $\sum_i P(B_i) = 1$:

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_j P(B_j)P(A|B_j)}$$

Posterior Distribution

After the experiment, the data has a certain likelihood for each parameter value $f_{X|\Theta}(x|\theta)$. The posterior distribution of θ is given by

$$\pi_{\Theta|X}(\theta|x) = \frac{\pi(\theta)L(\theta, x)}{\int_{-\infty}^{\infty} \pi(\theta)L(\theta, x)d\theta}$$

15 Bayesian Estimation

Marginal Distribution

If the distribution of θ has p.d.f. $\pi(\theta)$, then the marginal distribution of X has p.d.f.

$$f_X(x) = \int_{-\infty}^{\infty} \pi(\theta) f_{X|\Theta}(x|\theta)$$

Predictive Distribution

After the experiment, the distribution of a new data point Y is given by

$$f_{Y|X}(y|x) = \frac{\pi(\theta) f_{X|\Theta}(x|\theta) f_{X|\Theta}(y|\theta)}{\int_{-\infty}^{\infty} \pi(\theta) L(\theta, x) d\theta}$$

Question 46

An insurance company believes that claim sizes follow an inverse gamma distribution with $\alpha = 3$ and an unknown value of θ. This value of θ follows a gamma distribution with $\alpha = 5$ and $\theta = 1000$.

(a) Calculate the marginal distribution of the claim size.

(b) The company then observes the following sample of claims:

132	184	221	260	343	379	472	665	822	1,062
1,550	2,857	4,422							

What is the posterior distribution of θ?

(c) What is the predictive distribution of θ?

Loss Function

A loss function is a function $l_j(\hat{\theta}_j, \theta_j)$ is a measure of how much harm is done by obtaining an estimate of $\hat{\theta}_j$ when the true value is θ_j.
We then typically choose the estimate $\hat{\theta}_j$ to minimise the expected loss $\mathbb{E}(l_j(\hat{\theta}_j, \theta_j))$.

Examples

1. $l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$ is the squared-error loss. The estimator that minimises this expected loss is the posterior mean.
2. $l(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$ is the absolute-error loss. The estimator that minimises this expected loss is the posterior median.
3. $l(\hat{\theta}, \theta) = \begin{cases} 0 & \text{if } \hat{\theta} = \theta \\ 1 & \text{if } \hat{\theta} \neq \theta \end{cases}$ is the zero-one loss. The estimator that minimises this expected loss is the posterior mode.

Question 47

Recall from Question 46(b) that the posterior distribution of θ is a gamma distribution with $\alpha = 17$ and $\theta = 28.62476$. The predictive distribution had density function

$$f_{Y|X}(y|x) = C\frac{x^{17}}{(1000 + x)^{21}}$$

(a) Calculate the Bayes estimate for θ using a squared-error loss function.
(b) Calculate the expected value of the predictive distribution.
(c) Calculate the expected value of X based on the Bayes estimate for $\hat{\theta}$.

Question 48

An insurance company believes that the number of claims follows a Poisson distribution. It's prior distribution for the mean of the Poisson distribution is a Gamma distribution with $\alpha = 4$ and $\theta = 0.05$. It reviews 400,000 policies and finds that a total of 62,310 claims were made from these policies. Find a 95% credibility interval for λ.
(a) Using the HPD interval
(b) So that the probability of being above this interval and below this interval is equal.
(c) Using a normal approximation.

Definition

A prior distribution is a conjugate prior for a given model if the resulting posterior distribution is from the same family as the prior.

Example

For a Poisson distribution with parameter λ, a Gamma distribution is a conjugate prior, because the joint distribution is proportional to

$$
\begin{aligned}
&\lambda^{\alpha-1} e^{-\frac{\lambda}{\theta}} e^{-N\lambda} \lambda^{\sum X_i} \\
&= \lambda^{\alpha+\sum X_i - 1} e^{-\frac{\lambda}{\frac{1}{N+\frac{1}{\theta}}}} \\
&= \lambda^{\alpha+\sum X_i - 1} e^{-\frac{\lambda}{\frac{\theta}{N\theta+1}}}
\end{aligned}
$$

Which is the pdf of another Gamma distribution.

Question 49

Calculate the conjugate prior distribution for a distribution in the linear exponential family:

$$f_{X|\Theta}(x|\theta) = \frac{p(x)e^{r(\theta)x}}{q(\theta)}$$

Question 50

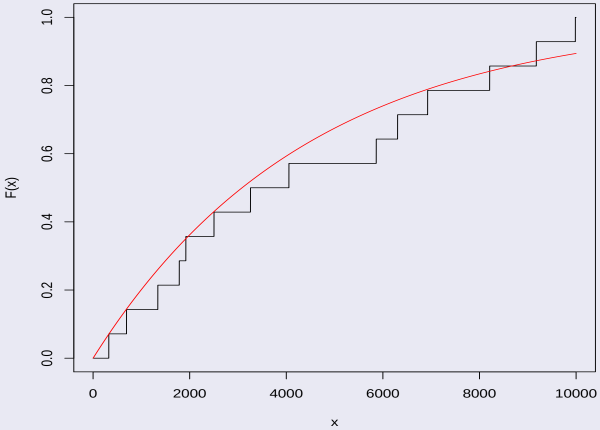
An insurance company is modeling claim severity. It collects the following data points:

325	692	1340	1784	1920	2503	3238	4054	5862
6304	6926	8210	9176	9984				

By graphically comparing distribution functions, assess the appropriateness of a Pareto distribution for modeling this data.

16.3 Graphical Comparison of Density and Distribution Functions

Answer to Question 50



Question 51

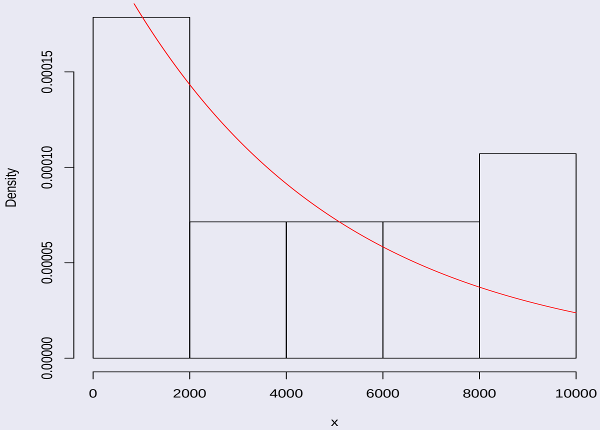
For the data from Question 50:

325	692	1340	1784	1920	2503	3238	4054	5862
6304	6926	8210	9176	9984				

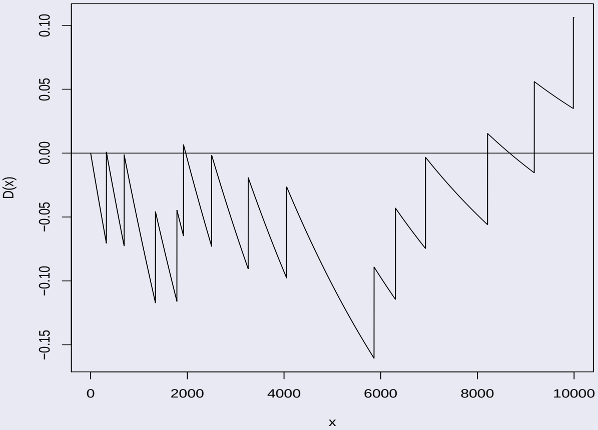
Graphically comparing density functions to assess the appropriateness of a Pareto distribution for modeling this data.

16.3 Graphical Comparison of Density and Distribution Functions

Answer to Question 51

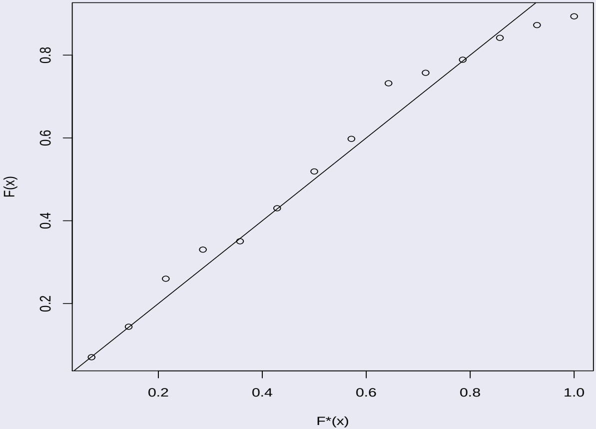


Answer to Question 51



16.3 Graphical Comparison of Density and Distribution Functions

Answer to Question 51



Hypothesis Tests

We test the following hypotheses:

H_0: The data came from a population with the given model.

H_1: The data did not come from a population with the given model.

16.4 Hypothesis Tests

Kolmogorov-Smirnov test

$$D = \max_{t \leqslant x \leqslant u} |F_n(x) - F(x)|$$

Anderson-Darling test

$$A^2 = n \int_t^u \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} f(x)\, dx$$

Chi-square Goodness-of-fit test

- Divide the range into separate regions, $t = c_0 < c_1 < \cdots < c_n = u$.
- Let O_i be the number of samples in the interval $[c_{i-1}, c_i)$.
- Let E_i be the expected number of sample in the interval $[c_{i-1}, c_i)$.

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Question 52

For the data from Question 50:

325	692	1340	1784	1920	2503	3238	4054	5862
6304	6926	8210	9176	9984				

Test the goodness of fit of the model using:
(a) The Kolmogorov-Smirnov test.
(b) The Anderson-Darling test.

Question 53

An insurance company records the following claim data:

Claim Amount	Frequency
0–5,000	742
5,000–10,000	1304
10,000–15,000	1022
15,000–20,000	830
20,000–25,000	211
More than 25,000	143

Use a Chi-square test to determine whether Claim size follows an exponential distribution.

Likelihood Ratio test

The Likelihood ratio test compares two nested models — $\mathcal{M}_0$ and $\mathcal{M}_1$.

Hypotheses

H_0: The simpler model describes the data as well as the more complicated model.

H_1: The more complicated model describes the data better than the simpler model.

We compute the parameters from both models by maximum likelihood. The test statistic is.

$$2(l_{\mathcal{M}_0}(x;\theta_0) - l_{\mathcal{M}_1}(x;\theta_1))$$

Under H_0, for large n, this follows a Chi-square distribution with degrees of freedom equal to the difference in number of parameters.

Question 54

An insurance company observes the following sample of claim data:

382	596	920	1241	1358	1822	2010	2417	2773
3002	3631	4120	4692	5123				

Use a likelihood ratio test to determine whether an exponential or a Weibull distribution fits this data better.

16.5 Selecting a Model

Comments on Model Selection

- Try to pick a model with as few parameters as possible. (Parsimony)
- Choice of model depends on the aspects that are important. Even if a formal test is used, the choice of which test depends on the aspects that are important.
- Aim is generalisability. The model should apply to future data. (Models which fit the given data well, but not new data are said to overfit.)
- Trying large numbers of models will lead to one which fits well just by chance.
- Experience is a valuable factor in deciding on a model.
- Sometimes knowledge of the underlying process may lead to a particular model (e.g. binomial).

Example

- An insurance company offers group life insurance to all 372 employees of a company.
- The premium is set at \$1,000 per year.
- The company notices that the average annual total claim over the past 7 years is \$126,000 — Far lower than the total premiums charged.

The company contacts the insurers and asks for a reduction in premiums on the basis that premiums are much larger than the average claim.

(a) Is this request reasonable?

(b) What would be a fair reduction in premium?

Definition

We assign full credibility to a policyholder's past history if we have sufficient data to use the policyholder's average claim for our premium estimate.

Criterion for Full Credibility

Let ξ be the (unknown) expected claim from a policyholder. We pick $r \geqslant 0$ and $0 < p < 1$. We assign full credibility to X if

$$P(|\overline{X} - \xi| < r\xi) > p$$

That is if with probability p, the relative error of $\overline{X}$ as an estimator for ξ is less than r.

Question 55

Recall our earlier example:

- An insurance company offers group life insurance to all 372 employees of a company.
- The premium is set at $1,000 per year.
- The average annual total claim over the past 7 years is $126,000.

Suppose that all policies have a death benefit of $98,000, and deaths of each employee are independent.
(a) Should the insurers assign full credibility to this experience? (Use $r = 0.05$ and $p = 0.95$.)
(b) How many years of past history are necessary to assign full credibility?

Question 56

A car insurance company is reviewing claims from a particular brand of car. It finds that over the past 3 years:

- it has issued 41,876 annual policies for this type of car.
- The average annual aggregate claim per policy is $962.14.
- The standard deviation of annual aggregate claim per policy is $3,605.52

(a) Should it assign full credibility to the historical data from this type of car?
(b) How many policies would it need in order to assign full credibility?

Question 57

Recall our original example:

- Group life insurance for 372 employees of a company.
- The premium is set at $1,000 per year.
- The average annual total claim over the past 7 years is $126,000.

All policies have a death benefit of $98,000, and deaths of each employee are independent.
In Question 55, we determined that this was not sufficient to assign full credibility to the data, and that 1191.034 years of claims data would be needed for full credibility.

> How much credibility should we assign to this data, and what should the resulting premium be?

Question 58

For a particular insurance policy, the average claim is $230, and the average claim frequency is 1.2 claims per year. A policyholder has enrolled in the policy for 10 years, and has made a total of 19 claims for a total of $5,822. Calculate the new premium for this policyholder if the standards for full credibility are:
(a) 421 claims for claim frequency, 1,240 claims for severity.
(b) 1146 claims for claim frequency, 611 claims for severity.
(c) 400 years for aggregate losses

17.5 Problems with this Approach

Problems

- No theoretical justification.
- Need to choose r and p arbitrarily.
- Doesn't take into account uncertainty in the book premium.

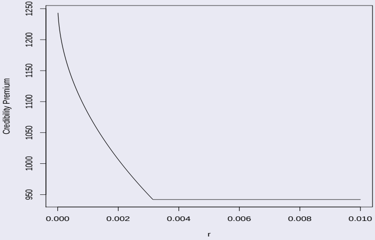
Question 59

An insurance company sells car insurance. The standard annual premium is $1,261. A car manufacturer claims that a certain model of its cars is safer than other cars and should receive a lower premium. The insurance company has issued 3,722 policies for this model of car. The total aggregate claims on these polices were $3,506,608. The variance of the annual aggregate claims on a policy is 8,240,268. Calculate the Credibility premium for different values of r and p.

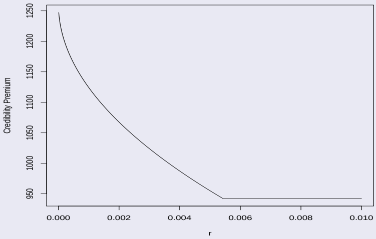
R-code for Question 59

```
#Limited Fluctuation Credibility Premium as r changes
p<-0.05
r<-(1:1000)/100000
Z<-sqrt(1221.565*r)/qnorm(1-p/2)
Z<-pmin(Z,1)
plot(r,Z*3506608/3722+(1-Z)*1261,type='l')
pdf("LFCredibilityChangeRPp=0.05.pdf")
plot(r,Z*3506608/3722+(1-Z)*1261,type='l',ylab="
   Credibility␣Premium")
dev.off()
```

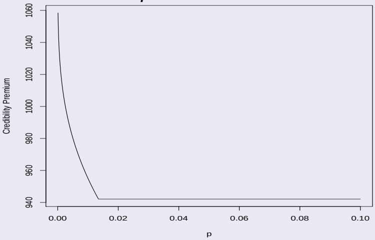
Answer to Question 59



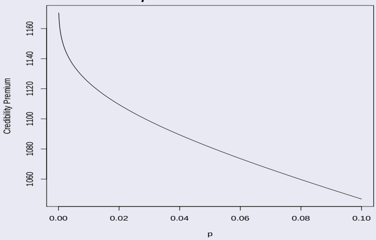
$p = 0.05$



$p = 0.01$



$r = 0.005$



$r = 0.001$

Assumptions

- Each policyholder has a risk parameter Θ, which is a random variable, but is assumed constant for that particular policyholder.
- Individual values of Θ can never be observed.
- The distribution of this risk parameter Θ has density (or mass) function $\pi(\theta)$, which is known. (We will denote the distribution function $\Pi(\theta)$.)
- For a given value $\Theta = \theta$, the conditional density (or mass) of the loss distribution $f_{X|\Theta}(x|\theta)$ is known.

18.2 Conditional Distributions and Expectation

Conditional Distributions (revision)

$$f_{X|\Theta}(x|\theta) = \frac{f_{X,\Theta}(x,\theta)}{\int f_{X,\Theta}(y,\theta)\,dy}$$
$$f_{X|\Theta}(x|\theta)f_\Theta(\theta) = f_{\Theta|X}(\theta|x)f_X(x)$$

Conditional Expectation (revision)

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|\Theta))$$
$$\mathrm{Var}(X) = \mathbb{E}(Var(X|\Theta)) + \mathrm{Var}(\mathbb{E}(X|\Theta))$$

Question 60

An insurance company models drivers as falling into two categories: frequent and infrequent. 75% of drivers fall into the frequent category. The number of claims made per year by a driver follows a Poisson distribution with parameter 0.4 for frequent drivers and 0.1 for infrequent drivers.

(a) Calculate the expectation and variance of the number of claims in a year for a randomly chosen driver.

(b) Calculate the expectation and variance of the number of claims in a year for a randomly chosen driver who made no claims in the previous year.

Question 61

The aggregate health claims (in a year) of an individual follows an inverse gamma distribution with $\alpha = 3$ and θ varying between individuals. The distribution of θ is a Gamma distribution with parameters $\alpha = 3$ and $\theta = 100$.
(a) Calculate the expected total health claims for a random individual.
(b) If an individual's aggregate claims in two consecutive years are \$112 and \$240, calculate the expected aggregate claims in the third year.

Question 62

The number of claims made by an individual in a year follows a Poisson distribution with parameter Λ. Λ varies between individuals, and follows a Gamma distribution with $\alpha = 4$ and $\theta = 0.1$.
(a) Calculate the expected number of claims for a new policyholder.
(b) Calculate the expected number of claims for a policyholder who has made *m* claims in the previous *n* years.

Problems with Bayesian Approach

- Difficult to Compute.
- Sensitive to exact model specification.
- Difficult to perform model selection for the unobserved risk parameter Θ.

Approach

- Credibility premium is a linear combination of book premium and personal history.

$$\alpha_0 + \sum_{i=1}^{n} \alpha_i X_i$$

- Coefficients are chosen to minimise Mean Squared Error (MSE)

$$\mathbb{E}\left(\mu(\Theta) - \left(\alpha_0 + \sum_{i=1}^{n} \alpha_i X_i\right)\right)^2$$

Question 63

Show that the solution which minimises the MSE satisfies:

$$\mathbb{E}(X_{n+1}) = \alpha_0 + \sum_{i=1}^{n} \alpha_i \mathbb{E}(X_i)$$

$$\mathrm{Cov}(X_i, X_{n+1}) = \sum_{j=1}^{n} \alpha_j \,\mathrm{Cov}(X_i, X_j)$$

Question 64

Suppose the X_i all have the same mean, the variance of X_i is σ^2, and the covariance $\text{Cov}(X_i, X_j) = \rho$. Calculate the credibility estimate for X_{n+1}.

Question 65

Suppose we have observations $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, which are the aggregate annual claims for each of two cars driven by an individual. We assume:

$$\begin{aligned}
\mathbb{E}(X_i) &= \mu \\
\mathbb{E}(Y_i) &= \nu \\
\mathrm{Var}(X_i) &= \sigma^2 \\
\mathrm{Var}(Y_i) &= \tau^2 \\
\mathrm{Cov}(X_i, X_j) &= \rho \quad \text{for } i \neq j \\
\mathrm{Cov}(Y_i, Y_j) &= \zeta \quad \text{for } i \neq j \\
\mathrm{Cov}(X_i, Y_j) &= \xi
\end{aligned}$$

Calculate the credibility estimate for $X_{n+1} + Y_{m+1}$.

18.5 The Buhlmann Model

Assumptions

- $X_1, \ldots, X_n$ are i.i.d. conditional on Θ.

We then define:

$\mu(\theta)=\mathbb{E}(X|\Theta = \theta)$ $\quad\quad$ $\mu=\mathbb{E}(\mu(\Theta))$

$\nu(\theta)=\mathrm{Var}(X|\Theta = \theta)$ $\quad\quad$ $\nu=\mathbb{E}(\nu(\Theta))$

$a=\mathrm{Var}(\mu(\Theta))$

Solution

$$\mathbb{E}(X_i) = \mu \qquad \mathrm{Var}(X_i) = \nu + a$$

$$\mathrm{Cov}(X_i, X_j) = a$$

Recall from Question 64, that the solution to this is:

$$\hat{\mu} = \frac{\left(\frac{v}{a}\right)}{n + \left(\frac{v}{a}\right)}\mu + \frac{n}{n + \left(\frac{v}{a}\right)}\overline{X}$$

Question 66

An insurance company offers group health insurance to an employer. Over the past 5 years, the insurance company has provided 851 policies to employees. The aggregate claims from these policies are \$121,336. The usual premium for such a policy is \$326. The variance of hypothetical means is 23,804, and the expected process variance is 84,036. Calculate the credibility premium for employees of this employer.

Question 67

An insurance company offers car insurance. One policyholder has been insured for 10 years, and during that time, the policyholder's aggregate claims have been \$3,224. The book premium for this policyholder is \$990. The expected process variance is 732403 and the variance of hypothetical means is 28822. Calculate the credibility premium for this driver next year.

18.6 The Buhlmann-Straub Model

Assumptions

- Each observation X_i (expressed as loss per exposure) has a (known) exposure m_i. The conditional variance of X_i is $\frac{v(\theta)}{m_i}$.

$$\text{Cov}(X_i, X_j) = a$$
$$\text{Var}(X_i) = \frac{v}{m_i} + a$$

Solution

$$\alpha_0 = \frac{\left(\frac{v}{a}\right)}{m + \frac{v}{a}}\mu \qquad \alpha_i = \frac{m_i}{m + \frac{v}{a}}$$

$$\hat{\mu} = \frac{\left(\frac{v}{a}\right)}{m + \frac{v}{a}}\mu + \frac{m}{m + \frac{v}{a}}\overline{X}$$

where $\overline{X}$ is the weighted mean $\sum_{i=1}^{n} \frac{m_i}{m} X_i$.

Question 68

For a group life insurance policy, the number of lives insured and the total aggregate claims for each of the past 5 years are shown in the following table:

Year	1	2	3	4	5
Lives insured	123	286	302	234	297
Agg. claims	0	\$300,000	\$200,000	\$200,000	\$300,000

The book rate for this policy premium is \$1,243 per life insured. The variance of hypothetical means is 120,384 and the expected process variance is 81,243,100. Calculate the credibility premium per life insured for the next year of the policy.

Question 69

A policyholder holds a landlord's insurance on a rental property. This policy is in effect while the property is rented out. The company has the following experience from this policy:

Year	1	2	3	4	5	6
Months rented	3	11	8	12	6	9
Agg. claims	0	\$10,000	0	0	\$4,000	0

The standard premium is \$600 per year for this policy. The variance of hypothetical means is 832076, and the expected process variance is 34280533 (both for annual claims). Calculate the credibility premium for the following year using the Buhlmann-Straub model.

Question 70

Show that if the Bayes premium is a linear function of X_i, and the expectation and variance of X are defined, then the Bayes premium is equal to the credibility premium.

Question 71

Show that if the model distribution is from the linear exponential family, and the prior is the conjugate prior, with $\frac{\pi(\theta_1)}{r'(\theta_1)} = \frac{\pi(\theta_0)}{r'(\theta_0)}$, where θ_0 and θ_1 are the upper and lower bounds for θ, then the Bayes premium is a linear function in X.

Approach

- Estimate the distribution of Θ from the data.
- Use this estimate to calculate the credibility estimate of μ.

Two possibilities

Either: We do not have a good model for the conditional or prior distribution. We only need the variances, so we estimate them non-parametrically.

or: We have a parametric model, such as a Poisson distribution, which allows us to estimate the variance more efficiently (assuming the model is accurate).

Question 72

An insurance company has the following aggregate claims data on a new type of insurance policy:

No.	Year 1	Year 2	Year 3	Year 4	Year 5	Mean	Variance
1	336	0	528	0	0	172.80	60595.2
2	180	234	0	2,642	302	671.60	1225822.8
3	0	0	528	361	0	177.80	62760.2
4	443	729	1,165	0	840	635.40	192962.3
5	0	0	0	0	0	0.00	0.0
6	196	482	254	303	0	247.00	30505.0
7	927	0	884	741	604	633.60	140653.7
8	0	601	105	130	327	232.60	56385.3

(a) Estimate the expected process variance and the variance of hypothetical means.
(b) Calculate the credibility premiums for each policyholder next year.

Question 73

An insurance company offers a group-life policy to 3 companies. These are the companies' exposures and aggregate claims (in millions) for the past 4 years:

Co.		Year 1	Year 2	Year 3	Year 4	Total
1	Exp	769	928	880	1,046	3,623
	Claims	1.3	1.5	0.8	1.7	5.3
2	Exp	1,430	1,207	949	1,322	4,908
	Claims	1.0	0.9	0.6	1.5	4.0
3	Exp	942	1,485	2,031	1,704	6,162
	Claims	1.1	1.4	1.9	2.0	6.4

Calculate the credibility premiums per life for each company in the fifth year.

Question 74

In a particular year, an insurance company observes the following claim frequencies:

No. of Claims	Frequency
0	3951
1	1406
2	740
3	97
4	13
5	3

Assuming the number of claims an individual makes follows a Poisson distribution, calculate the credibility estimate for number of claims for an individual who has made 6 claims in the past 3 years.

Question 75

Assume annual claims from one policyholder follow a Poisson distribution with mean Λ. The last 4 years of claims data are:

Claims	0	1	2	3	4	5	6	7	8	9
1 year	3951	1406	740	97	13	3	0	0	0	0
2 years	3628	2807	1023	461	104	13	4	0	1	0
3 years	2967	4032	2214	890	734	215	131	22	0	2
4 years	1460	2828	2204	985	747	358	194	43	8	0

Calculate the credibility estimate of Λ for an individual who made 2 claims in the last 3 years of coverage.

20 Simulation

Problem

We have some distribution which can be specified in a complicated way, making it hard to describe.

Solution

Generate data following this distribution and study this sample.

Examples

- Bootstrapping
- Calculating aggregate losses on a portfolio of insurance policies with deductibles and policy limits
- Incorporating the effect of interest rates into the net profit or loss on life insurance policies.
- Modelling the effect of major events on aggregate losses (e.g. earthquakes, hurricanes, etc.)

Psuedorandom Numbers

Definition

A pseudorandom number is a number generated by a formula, such that to someone who does not know the formula, the number is indistinguishable from a random number following a specified distribution.

Notes

- Computers typically only include formulae to generate pseudo-random numbers from a uniform distribution.
- Random numbers following other distributions can be generated by inversion.

Question 76

A computers random number generator provides the following three numbers from a uniform distribution:

0.1850620 0.8613517 0.3607076

Use these samples to generate three random numbers following:
(a) A normal distribution with $\mu = 2$ and $\sigma^2 = 9$.
(b) A Pareto distribution with $\alpha = 4$ and $\theta = 2400$.
(c) A Poisson distribution with $\lambda = 2.4$.

Question 77

An insurance company classifies drivers as *good*, *average* or *bad*. For each type, the distribution of the loss amount for an accident is a Pareto distribution, whose parameters are given in the following table:

Type	Proportion of drivers	α	θ
Good	0.02	5	2,800
Average	0.86	4	4,000
Bad	0.12	3	4,200

They simulate the following uniform random numbers:

0.29351756 0.11768610 0.47362823 0.13843535

Use these to simulate:
(a) two loss amounts for two different drivers
(b) two loss amounts for the same driver

Question 78

A group life insurance policy has three possible decrements: death, disability and withdrawl. The probabilites of these events occuring in a year are 0.01, 0.02 and 0.12 respectively. The insurance company wants to simulate the number of each decrement from 720 policies. They simulate the following uniform random variables.

0.3876723	0.2534800	0.2954348	0.6049291

Use these to generate a simulated number of each decrement.

Question 79

An insurance policy has exits only through death or lapses. The probability of death in the first year is 0.01. The probability of lapse in the first year is 0.02. The probability of death in the second year is 0.015, the probability of lapse in the second year is 0.04, and the probability of death in the third year is 0.02. From the simulated numbers

0.8579075	0.8193713	0.4031135	0.7313493
0.9613431	0.7735622	0.9745215	0.6261118

calculate the simulated number of individuals from 200 policies who die during the third year.

Simulating from $(a, b, 0)$-class distributions.

- Simulate times between events from an exponential distribution with parameter λ_k, where k is the number of events already observed.
- Set $\lambda_k = c + dk$, where c and d depend on the distribution to be simulated.
- Simulate until the total time exceeds 1.

Question 80

Show that the values of c and d are as given in the following table:

Distribution	c	d
Poisson (λ)	λ	0
Binomial (n, p)	$n\log(1-p)$	$-\log(1-p)$
Negative Binomial (r, β)	$r\log(1+\beta)$	$\log(1+\beta)$

Question 81

You generate the following sample from a uniform distribution:

0.9587058	0.4975469	0.7957639	0.1762183
0.8649957	0.4639014	0.4426729	0.4197114
0.4212635	0.3984598	0.4043391	0.3122119

Use the stochastic process method above, and these random numbers to generate:
(a) A binomial random variable with $n = 20$ and $p = 0.14$.
(b) A Poisson random variable with $\lambda = 6$.
(c) A negative binomial random variable with $r = 3$ and $\beta = 2$.

20.2.4 Generating Normal Random Variables

Problem with Inversion

The distribution function is calculated by numerical integration. The errors in this are cumulative, so become relatively large near the tails.

Box-Muller Transform

- Generate U_1 and U_2, independent uniform random variables.
- $Z_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$ and $Z_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$

Polar Method

- Generate U_1 and U_2, independent uniform random variables.
- Calculate $X_1 = 2U_1 - 1, X_2 = 2U_2 - 1$
- If $W = {X_1}^2 + {X_2}^2 \geqslant 1$, discard this sample.
- Let $Y = \sqrt{\frac{-2\log(W)}{W}}$
- Calculate $Z_1 = X_1 Y$ and $Z_2 = X_2 Y$.

Question 82

You generate the following sample from a uniform distribution:

0.9974532	0.4429451	0.6159707	0.6626078

Use these random numbers to generate two normal random variables using
(a) A Box-Muller transform.
(b) The polar method.

Question 83

An insurance company wants to calculate the probability of a loss exceeding $200,000. It wants the error in its estimated probability to be at most 0.0001. How many simulations does it need to perform to achieve this accuracy?

Question 84

An insurance company estimates that losses on a certain group of policies follow a Pareto distribution with $\alpha = 2.5$ and $\theta = 4,300$. The policies have a deductible of $1,000 and a policy limit of $1,000,000. The number of losses in a year follows a negative binomial distribution with $r = 12$ and $\beta = 1.6$.
Use a simulation to estimate the 95th percentile of the aggregate loss distribustion, stopping when the estimate has a 95% chance of being within $100 of the true value.

R-code for Question 84

```
#Simulate losses
U<-runif(1000000000)
paret<-4300*((1-U)^(-0.4)-1)
dim(paret)<-c(10000000,100)

#Simulate no. of losses
U<-runif(1000000000)
expon<-log(1-U)
dim(expon)<-c(10000000,100)
lambda<-rep(1,10000000)%*%t((12:111)*log(2.6))
expon<--log(1-U)
expon<-expon/lambda
exponSum<-apply(expon,1,cumsum)
```

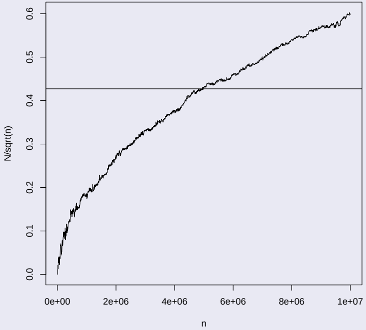
R-code for Question 84

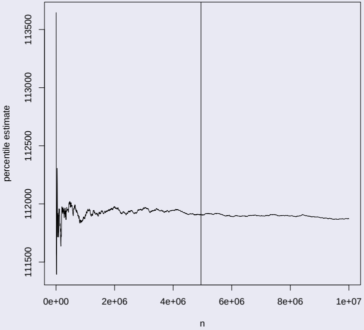
```
#calculate aggregate losses
ag<-rowSums(paret*(t(exponSum)<1))

#Calculate percentile
rnf<-rep(0,2000)
for(i in 1:2000){
rnf[i]<-quantile(ag[1:(5000*i)],0.95)
}

#Count points between pi_0.95 and (pi_0.95)+100
N<-rep(0,2000)
for(i in 1:2000){
N[i]<-sum(ag[which(ag[1:(5000*i)]>rnf[i])]<(rnf[i]+100))
}
```

Answer to Question 84





Applications of Simulation

- Calculating aggregate losses.
- Calculating risk estimates
- Calculating *p*-values
- Calculating MSE, confidence intervals, etc.

Question 85

An insurance company has 3482 policies. The policies have deductibles of \$500, \$1,000 or \$5,000 with probabilities 0.5, 0.35, and 0.15 respectively. The policies all have limit \$1,000,000. Loss amounts follow a Pareto distribution with $\alpha = 2.2$ and $\theta = 6,000$. The number of losses on a given policy follows a zero-modified ETNB distribution with $r = -0.6$ and $\beta = 3.1$ and $p_0 = 0.91$. The insurance company has taken out stop-loss insurance which pays 90% of aggregate claims above \$10,000,000,000.
Use simulation to find the expected aggregate loss on these policies.

Question 86

Use simulation to estimate the TVaR of a Gamma distribution with $\alpha = 13$ and $\theta = 1000$.

R-code for Question 86

```
#Use simulation to estimate the TVaR of a gamma
#distribution with alpha=13 and theta=1000

U<-runif(1000000)
X<-qgamma(U,13,scale=1000)

VaR<-quantile(X,0.95)
tvar<-mean(X[X>VaR])
#We could do more sophisticated techniques to
#improve the answer, but for large enough sample
#size, this should be sufficient.
```

Question 87

Testing whether it is appropriate to model a particular dataset using a Gamma distribution, the fitted parameters are $\alpha = 3.7$ and $\theta = 1,352$ and the Anderson-Darling test statistic is 1.84, based on sample size 186. Use a simulation to estimate the *p*-value of this statistic.

R-code for Question 87

```
#estimates the p-value of the test statistic for the
    Anderson-Darling test.
library(stats4)

U<-runif(1860000)
X<-qgamma(U,3.7,scale=1352)

dim(X)<-c(10000,186)

negll<-function(al,th){
    return(186*al*log(th)+sum(X[i,])/th+186*log(gamma(al)
        )-(al-1)*sum(log(X[i,])))
}

st<-list(al=3.7,th=1352)
```

R-code for Question 87

```
ests<-rep(0,20000)
dim(ests)<-c(10000,2)
for(i in 1:10000){
    ests[i,]<-attr(mle(negll,st),"coef")
}

AD<-rep(0,10000)
for(i in 1:10000){
    Y<-pgamma(sort(X[i,]),ests[i,1],,ests[i,2])
    AD[i]<-186*(sum((1-(1:185)/186)^2*(log(1-Y[1:185])-
        log(1-Y[2:186])))+sum(((1:186)/186)^2*(log(c(Y
        [2:186],1))-log(Y)-log(1-Y[1])-1)
}

sum(AD>1.84)
```