Instituto Superior de Economia e Gestão

UNIVERSIDADE TÉCNICA DE LISBOA

MASTER IN ACTUARIAL SCIENCE

Risk Models

27/01/2012

Time allowed: 3 hours

Instructions:

1. This paper contains 9 questions and comprises 4 pages including the title page.
2. Enter all requested details on the cover sheet.
3. You have 10 minutes reading time. You must not start writing your answers until instructed to do so.
4. Number the pages of the paper where you are going to write your answers.
5. Attempt all 9 questions.
6. Begin your answer to each of the 9 questions on a new page.
7. Marks are shown in brackets. Total marks: 200.
8. Show calculations where appropriate.
9. An approved calculator may be used.
10. The distributed formulary and the Formulae and Tables for Actuarial Examinations (the 2002 edition) may be used. Note that the parametrization used for the different distributions is that of the distributed formulary.

1. For a portfolio of policies, you are given:
   (i) There is no deductible and the policy limit varies by policy.
   (ii) A sample of ten claims is:

   350 350 500 500 500+ 1000 1000+ 1000+ 1200 1500

   where the symbol + indicates that the loss exceeds the policy limit (i.e. 500+ means that the policy limit is 500 and the claim exceeds 500)
   (iii) $S_n(1250)$ is the Kaplan-Meier estimate of $S(1250)$
   (iv) $\hat{S}(1250)$ is the maximum likelihood estimate of *S*(1250) under the assumption that the losses follow an exponential distribution.

   a) Get a confidence interval for $S(1250)$ based on the log transformed method and the Kaplan-Meier estimator. [marks 15]
   b) Get another confidence interval for $S(1250)$ based on the maximum likelihood estimator. [marks 15]
   c) Did you use asymptotic results to answer to the previous questions? Comment about the use of such results in this problem. [marks 5]

2. For a survival study, you are given:
   (i) Deaths occurred at times $y_1, y_2, \cdots, y_9$.
   (ii) The Nelson-Aalen estimates of the cumulative hazard function at y_3 and y_4 are $\hat{H}(y_3) = 0.4128$ and $\hat{H}(y_4) = 0.5691$
   (iii) The estimated variances of the estimators in (ii) are $\hat{\mathrm{var}}\left(\hat{H}(y_3)\right) = 0.009565$ and $\hat{\mathrm{var}}\left(\hat{H}(y_4)\right) = 0.014448$

   Determine the number of deaths at y_4. [marks 15]

3. You study five lives to estimate the time from the onset of a disease to death. The times to death are: 2; 3; 3; 3; 7. Using a triangular kernel,

$$k_y(x) = \begin{cases} (x - y + b)/b^2 & y - b \le x \le y \\ (y + b - x)/b^2 & y \le x \le y + b \end{cases}$$

   with bandwidth 2, estimate the density function at 2.5. [marks 20]

4. You are given the following claim data for automobile policies:

   200 255 295 320 360 420 440 490 500 520 1020

Assume that the claims follow an exponential distribution. Using the percentile method (45th percentile), obtain an estimate for the mean of the distribution and get an estimate for the 95th percentile of the distribution. [marks 15]

5. Let $f(x|\theta)$ be the density function (and $F(x|\theta)$ the distribution function) of the claim amounts for a given risk. Assume that the payments are made in excess of the deductible (if there is a deductible). Our purpose is to estimate θ, the parameter of the claim amount distribution. You are given the following information about a group of policies:

Deductible	No	No	10	25	No	No
Payment	5	15	60	100	500	500
Policy limit	50	50	100	100	500	1000

Write the likelihood function. [marks 10]

6. Let the density function of the random variable X be given by $f(x|\theta)=(1-\theta)^{x}\theta$, $x=0,1,\cdots$, and $0<\theta<1$. However, due to sampling problems we are not able to observe the 0 values.

Observed values	1	2	3	4	5
Frequency	101	26	14	2	1

a) Obtain a maximum likelihood estimate for θ. [marks 15]

b) Determine a 95% approximate confidence interval for $\Pr(X=0)$.
Determine a 95% approximate confidence interval for $\Pr(X=1)$. [marks 15]

7. Assume that X, given θ, follows an exponential distribution with mean $1/\theta$. From a Bayesian point of view we define that the prior for θ is an exponential distribution with mean 0.1.

a) Assuming that we observed the sample $(x_1,x_2,\cdots,x_n)$, show that the posterior for θ is a gamma distribution with parameters $n+1$ and $\left(10+\sum_{i=1}^{n}x_i\right)^{-1}$ [marks 10]

b) If the observed sample was (4.0; 5.0; 6.0), what is the probability that θ is greater than 0.1. [marks 10]

c) Show that the predictive distribution of y, given the sample (4.0; 5.0; 6.0), is a Pareto distribution with parameters 4 and 25. [marks 15]

8. You are given:
   (i) A random sample of 100 losses from a Weibull distribution gives the following statistics: $\overline{x} = 652.6836$; $s = 281.0159$; $\min x_i = 107.6066$; $\max x_i = 1418.844$; $\sum_{i=1}^{n} x_i = 65268.36$; $\sum_{i=1}^{n} x_i^2 = 50417618$; $\sum_{i=1}^{n} \ln x_i = 637.44753$
   (ii) At the maximum likelihood estimates of θ and τ, $\sum_{i=1}^{n} \ln(f(x_i \mid \hat{\theta}, \hat{\tau})) = -702.6301$
   (iii) When $\tau = 2$, the maximum likelihood estimate of θ is 710.0533.

   Use the likelihood ratio test to test the hypothesis $H_0 : \tau = 2$ against $H_1 : \tau \neq 2$ and conclude. [marks 20]

9. An actuary needs to estimate the dispersion of claims population. However he knows from past experience that this population has heavy tails and is highly skewed and then the usual measure, the sample standard deviation, can be influenced by extreme observations. To overcome this situation, he decides to use a more robust measure, the median absolute deviation from the median (MAD), which is defined as follows. If we consider a sample of size n, the MAD is the median of the numbers $y_i = |x_i - \tilde{x}|$ where $\tilde{x}$ is the median of the sample. Let us consider that we observed a sample of size 15. Explain how to use the bootstrap to approximate the MAD sampling distribution. [marks 20]

SOLUTION

1.

a)

j	y_j	s_j	r_j	$\frac{(r_j - s_j)}{r_j}$	$\prod_{i=1}^{j} \frac{(r_i - s_i)}{r_i}$
1	350	2	10	0.8	0.8
2	500	2	8	0.75	0.6
3	1000	1	5	0.8	0.48
4	1200	1	2	0.5	0.24
5	1500	1	1	0	0

$$S_n(1250) = 0.24$$

$$\hat{\text{var}}\, S_n(1250) \approx 0.24^2 \left(\frac{2}{80} + \frac{2}{48} + \frac{1}{20} + \frac{1}{2} \right) = 0.03552$$

$$U = \exp\left(1.96 \frac{\sqrt{0.03552}}{0.24 \times \ln 0.24} \right) = 0.3401$$

The 95% confidence interval is $\left(0.24^{1/0.3401}; 0.24^{0.3401}\right)$ i.e. $\left(0.0151; 0.6155\right)$

b) Uncensored observations: $\ell_i(\theta) = -\ln\theta - x_i / \theta$ where x_i is the observed value

Censored observations: $\ell_i(\theta) = -u_i / \theta$ where u_i is the censoring point

$$\ell(\theta) = -7\ln\theta - \frac{\sum x_i + \sum u_i}{\theta} = -7\ln\theta - \frac{7900}{\theta}$$

$$\ell'(\theta) = -\frac{7}{\theta} + \frac{7900}{\theta^2} \qquad \ell'(\hat{\theta}) = 0 \Leftrightarrow \hat{\theta} = \frac{7900}{7} = 1128.571$$

$$\ell''(\theta) = \frac{7}{\theta^2} - \frac{15800}{\theta^3} \qquad \ell''(\hat{\theta}) = -5.49591 \times 10^{-6}$$

$$\hat{S}(1250) = e^{-1250/1128.581} = 0.3304$$

Solution 1: determine a confidence interval for θ and then the confidence interval for $S(1250) = e^{-1250/\theta}$

$$\hat{\text{var}}(\hat{\theta}) = -1/\ell''(\hat{\theta}) = 181953.3528$$

Conf. interval for θ: $0.3304 \pm 1.96\sqrt{181953.3528}$ →(292.514; 1694.629)

Conf. interval for $S(1250) = e^{-1250/\theta}$ →(0.0211; 0.5630)

Solution 2: Use the delta method to approximate the variance of $\hat{S}(1250) = e^{-1250/\hat{\theta}}$ and then determine the confidence interval for $S(1250)$

$g(\theta) = e^{-1250/\theta}$; $g'(\theta) = (1250/\theta^2)\, e^{-1250/\theta}$;

$$\hat{\text{var}}(g(\hat{\theta})) = \left(g'(\hat{\theta})\right)^2 \hat{\text{var}}(\hat{\theta}) = 0.019126$$

Conf. interval for $S(1250) = e^{-1250/\theta}$ →(0.0593; 0.6014)

c) Yes we used asymptotic results both in question a) – approximation to the normal distribution for instance – and in question b) – asymptotic distribution of the maximum likelihood estimators. The use of asymptotic results is not appropriate since it is not reasonable to consider that the sample size is large.

2. As $\sum_{i=1}^{3} \frac{s_i}{r_i} = 0.4128$ and $\sum_{i=1}^{4} \frac{s_i}{r_i} = 0.5691$ we get $\frac{s_4}{r_4} = 0.1563$.

As $\sum_{i=1}^{3} \frac{s_i}{r_i^2} = 0.009565$ and $\sum_{i=1}^{4} \frac{s_i}{r_i^2} = 0.014448$ we get $\frac{s_4}{r_4^2} = 0.004883$.

We must solve the following system in order to get s_4:

$$\begin{cases} s_4 = 0.1563\times r_4 \\ s_4 = 0.004883\times r_4^2 \end{cases} \Leftrightarrow \begin{cases} s_4 = 0.1563\times r_4 \\ 0.1563\times r_4 = 0.004883\times r_4^2 \end{cases} \Leftrightarrow \begin{cases} s_4 = 0.1563\times r_4 \\ r_4 = \dfrac{0.1563}{0.004883} \end{cases} \Leftrightarrow \begin{cases} s_4 = \dfrac{0.1563^2}{0.004883} \\ r_4 = \dfrac{0.1563}{0.004883} \end{cases}$$

Then $s_4 = 5$.

3. You study five lives to estimate the time from the onset of a disease to death. The times to death are: 2; 3; 3; 3; 7. Using a triangular kernel

$$k_y(x) = \begin{cases} (x-y+b)/b^2 & y-b \le x \le y \\ (y+b-x)/b^2 & y \le x \le y+b \end{cases}$$

with bandwidth 2, estimate the density function at 2.5.

At $x = 2.5$ the only sample values y_j for which $k_{y_j}(2.5) > 0$ are $y_1 = 2$ and $y_2 = 3$.

Then

$$\hat{f}(2.5) = p(2)\times k_2(2.5) + p(3)\times k_3(2.5) = 0.2\times\frac{2+2-2.5}{4} + 0.6\times\frac{2.5-3+2}{4} = 0.3$$

4. $\hat{\pi}_g = (1-h)\,x_{(j)} + h\,x_{(j+1)}$ where $j = \lfloor (n+1)\,g \rfloor$ and $h = (n+1)\,g - j$

$g = 0.45$; $n = 11$; $j = \lfloor (n+1)\,g \rfloor = 5$; $h = (n+1)\,g - j = 0.4$;

$$\hat{\pi}_{0.45} = 0.6\times x_{(5)} + 0.4\times x_{(6)} = 384$$

$$0.45 = 1 - e^{-\pi_{0.45}/\theta} \Leftrightarrow e^{-\pi_{0.45}/\theta} = 0.55 \Leftrightarrow \frac{\pi_{0.45}}{\theta} = -\ln 0.55 \Leftrightarrow \theta = -\frac{\pi_{0.45}}{\ln 0.55}$$

Then $\tilde{\theta} = -384/\ln 0.55 = 642.316$

Now we will estimate the 95th percentile of the distribution,

$\tilde{\pi}_{95} = -\tilde{\theta}\ln 0.05 = 1924.205$

5. $L(\theta) = f(5\,|\,\theta)\times f(15\,|\,\theta)\times\dfrac{f(70\,|\,\theta)}{1-F(10\,|\,\theta)}\times\dfrac{1-F(125\,|\,\theta)}{1-F(25\,|\,\theta)}\times\big(1-F(500\,|\,\theta)\big)\times f(500\,|\,\theta)$

6.

a) $L(\theta) = \prod_{i=1}^{n} \dfrac{f(x_i\,|\,\theta)}{1-\Pr(X=0\,|\,\theta)} = \prod_{i=1}^{n} \dfrac{(1-\theta)^{x_i}\,\theta}{1-\theta} = \prod_{i=1}^{n} (1-\theta)^{x_i-1}\,\theta$

$$\ell(\theta)=\sum_{i=1}^{n}\left(\left(x_i-1\right)\ln\left(1-\theta\right)+\ln\theta\right)$$

$$\ell'(\theta)=\sum_{i=1}^{n}\left(-\frac{\left(x_i-1\right)}{\left(1-\theta\right)}+\frac{1}{\theta}\right)=-\frac{n\,\overline{x}-n}{1-\theta}+\frac{n}{\theta}$$

$$\ell'(\theta)=0\Leftrightarrow\frac{n\,\overline{x}-n}{1-\theta}=\frac{n}{\theta}\Leftrightarrow\frac{1-\theta}{\theta}=\overline{x}-1\Leftrightarrow\frac{1}{\theta}=\overline{x}\Leftrightarrow\theta=\frac{1}{\overline{x}}$$

$$\ell''(\theta)=-n\left(\frac{\overline{x}-1}{(1-\theta)^2}+\frac{1}{\theta^2}\right)<0$$

Then the maximum likelihood estimate for θ is $\hat{\theta}=1/\overline{x}=0.6923$

b) $\Pr(X=0\,|\,\theta)=\theta$; $\overset{\wedge}{\Pr}(X=0)=\hat{\theta}$; $\overset{\wedge}{\operatorname{var}}\left(\overset{\wedge}{\Pr}(X=0)\right)=\overset{\wedge}{\operatorname{var}}(\hat{\theta})\approx-1/\ell''(\hat{\theta})=0.001024$

The 95% confidence interval is then $0.6923\pm1.96\sqrt{0.001024}$ → (0.6296; 0.7550)

$\Pr(X=1\,|\,\theta)=\theta(1-\theta)$; $\overset{\wedge}{\Pr}(X=0)=\hat{\theta}(1-\hat{\theta})=0.2130$; $g(\theta)=\theta(1-\theta)$; $g'(\theta)=1-2\theta$

$\overset{\wedge}{\operatorname{var}}\left(\overset{\wedge}{\Pr}(X=1)\right)=\overset{\wedge}{\operatorname{var}}(\hat{\theta}(1-\hat{\theta}))\approx\left(g'(\hat{\theta})\right)^2\left(-1/\ell''(\hat{\theta})\right)=.147917\times0.001024=0.000151$

The 95% confidence interval is then $0.2130\pm1.96\sqrt{0.000151}$ →(0.1889; 0.2371)

7. Assume that *X*, given θ, follows an exponential distribution with mean $1/\theta$. From a Bayesian point of view we define that the prior for θ is an exponential distribution with mean 0.1.

a) $f(x\,|\,\theta)=\theta e^{-\theta x}$, $x>0$, $\theta>0$ $\qquad L(\theta\,|\,\mathbf{x})=\theta^n e^{-\theta t}$ with $t=\sum_{i=1}^{n}x_i$

$\pi(\theta)=(1/0.1)e^{-\theta/0.1}=10e^{-10\theta}$, $\theta>0$

$\pi(\theta\,|\,\mathrm{x})\propto10e^{-10\theta}\times\theta^n e^{-\theta t}\propto\theta^n e^{-\theta(t+10)}$ i.e. the posterior is a gamma distribution with parameters $n+1$ and $1/\left(10+\sum_{i=1}^{n}x_i\right)$

b) Observed sample: (4.0; 5.0; 6.0) then $n=3$ and $\sum_{i=1}^{n}x_i=15$.

$\Pr(\theta>0.1\,|\,\mathbf{x})=?$

$\theta\,|\,\mathbf{x}\sim G\left(4;1/25\right)$ then $2\times25\times\theta\,|\,\mathbf{x}\sim\chi^2_{(8)}$

$\Pr(\theta>0.1\,|\,\mathbf{x})=\Pr(50\theta>5.0\,|\,\mathbf{x})=0.7576$

c) Show that the predictive distribution of y, given the sample (4.0; 5.0; 6.0) is a Pareto distribution with parameters 4 and 25.

$$f_{Y|\mathbf{X}}(y\,|\,\mathbf{x})=\int_0^\infty\theta e^{-\theta y}\frac{\theta^3e^{-25\theta}}{\left(1/25\right)^4\Gamma(4)}d\theta=\frac{25^4}{\Gamma(4)}\int_0^\infty\theta^4e^{-(y+25)\theta}\,d\theta$$

$$=\frac{25^4}{\Gamma(4)}\frac{\Gamma(5)}{(y+25)^5}\int_0^\infty\frac{(y+25)^5}{\Gamma(5)}\theta^4e^{-(y+25)\theta}\,d\theta$$

$$=\frac{4\times25^4}{(y+25)^5}\qquad y>0$$

8. $H_0: \tau = 2$ against $H_1: \tau \neq 2$

$\lambda_{obs} = \dfrac{L(\hat{\hat{\theta}} \mid \tau = 2)}{L(\hat{\theta}, \hat{\tau})}$ where $\hat{\theta}$ and $\hat{\tau}$ are the maximum likelihood estimates of θ and τ respectively and $\hat{\hat{\theta}}$ is the maximum likelihood estimate of θ when $\tau = 2$.

Then $\ln \lambda_{obs} = \ln L(\hat{\hat{\theta}} \mid \tau = 2) - \ln L(\hat{\theta}, \hat{\tau})$

The critical value $\alpha = 0.05$ is given by a χ^2 distribution with 1 degree of freedom →3.841

$\ln L(\hat{\theta}, \hat{\tau}) = -702.6301$

$\ln L(\hat{\hat{\theta}} \mid \tau = 2) = \ln L(710.0533, 2) = \sum_{i=1}^{n} \left(\ln \tau - \tau \ln \theta + (\tau - 1) \ln x_i - (x_i / \theta)^{\tau} \right)$

$= 100 \times \left(\ln 2 - 2 \ln 710.0533 \right) + \sum_{i=1}^{n} \ln x_i - (1/\theta^2) \sum_{i=1}^{n} x_i^2$

$\ln L(\hat{\hat{\theta}} \mid \tau = 2) = -706.3059$

The observed value of the test statistic is then

$-2 \times \ln \lambda_{obs} = -2 \times \left(-706.3059 + 702.6301 \right) = 7.3515$

Conclusion: We reject the null hypothesis.

9. How to use the bootstrap to approximate the MAD sampling distribution (The first step is not mandatory)

i. Calculate the MAD of the observed sample, MAD^*. We sort the sample – $\left(x_{(1)}, x_{(2)}, \cdots, x_{(15)} \right)$ is the sorted sample – and pick $\tilde{x} = x_{(8)}$. Then we calculate $y_i = | x_i - \tilde{x} |$ and we sort them → $\left(y_{(1)}, y_{(2)}, \cdots, y_{(15)} \right)$. Then we get $MAD^* = y_{(8)}$

ii. Define B as the number of bootstrap samples to be generated.

iii. For each bootstrap generation, $j = 1, 2, \cdots, B$ do:
   a. Resample (with replacement) from the original sample;
   b. Calculate the MAD of the bootstrap sample, MAD_j, using the same procedure as before;

iv. Now, using the B bootstrap observations we can approximate the sampling distribution by means of a histogram or using the empirical cumulative distribution function.