MTH5126 - Statistics for Insurance

Academic Year: 2022-23 **Semester: B**

Worksheet 5 - Solutions

Q1. Extreme value theory

The claim amounts in a general insurance portfolio are independent and follow an exponential distribution with mean £2,500.

(i) Calculate the probability that an individual claim will exceed £10,000.

(ii) Calculate the probability that, in a sample of 100 claims, the largest claim will exceed £10,000 using:

(a) an exact method

(b) an approximation based on a Gumbel-type GEV distribution.

You are given that, for an exponential distribution with parameter λ, the approximate distribution of $\max\{X_1, X_2, \ldots, X_n\}$ for large n is a Gumbel-type GEV distribution with CDF:

$H(x) = \exp\{ -\exp[-(\frac{x-\alpha_n}{\beta_n})]\}$

where $\alpha_n = \frac{1}{\lambda}\ln n$ and $\beta_n = \frac{1}{\lambda}$

(iii) State the 2 key assumptions made in (ii) (a).

Answer:

(i)

$P(X > 10\,000) = 1 - P(X \leq 10\,000)$

Now:

$X \sim Exp(1/2500)$

$F_X(x) = 1 - \exp(-1/2500 * x)$

$F_X(10\,000) = 1 - \exp(-1/2500 * 10\,000) = 1 - \exp(-4)$

So:

$P(X > 10\,000) = 1 - P(X \leq 10\,000) = 1 - [1 - \exp(-4)] = \exp(-4) = 0.018316$

(ii) (a) Using an exact method

$P(X_M > 10\,000)$, where $X_M = \max\{ X_1, X_2, \ldots, X_{100}\}$

$= 1 - P(X_M \leq 10\,000)$

$= 1 - P(X_1 \leq 10\,000, X_2 \leq 10\,000, \ldots, X_{100} \leq 10\,000)$

$= 1 - P(X_1 \leq 10\,000)\, P(X_2 \leq 10\,000) \ldots P(X_{100} \leq 10\,000)$, because X_i's are independent

$= 1 - [P(X \leq 10\,000)]^{100}$, because X_i's are identical

$= 1 - [F(10\,000)]^{100}$

$= 1 - [1 - \exp(-1/2500 * 10000)]^{100}$

$= 1 - [1 - 0.018316]^{100}$

$= 0.8425$

(ii) (b) Using an approximate method

The approximate distribution of $X_M = \max\{ X_1, X_2, \ldots, X_{100}\}$ is a Gumbel-type GEV distribution with CDF:

$H(x) = \exp\{ -\exp[-(\frac{x - \alpha_{100}}{\beta_{100}})]\}$

where $\alpha_{100} = 2500 \ln 100$ and $\beta_{100} = 2500$

$P(X_M > 10\,000)$, where $X_M = \max\{ X_1, X_2, \ldots, X_{100}\}$

$= 1 - P(X_M \leq 10\,000)$

$\approx 1 - \exp\{ -\exp[-(\frac{10000 - \alpha_{100}}{\beta_{100}})]\}$

$= 1 - \exp\{ -\exp[-(\frac{10000 - 2500 \ln 100}{2500})]\}$

$= 1 - 0.1602 = 0.8398$

(iii) 2 key assumptions made in (ii)(a):

- Claim amounts are independent

- All claims follow an exponential distribution with mean £2,500

Q2.

(i) Explain why Extreme Value Theory (EVT) models can be useful. [2]

A sports scientist is interested in analysing the probability that the javelin world record may be broken next year and is intending to use EVT to do this. The sports scientist has obtained data for the distances of all javelin throws from all javelin competitions last year. The total number of throws recorded was 3,000. The sports scientist has carried out an EVT analysis using the Generalised Pareto Distribution by selecting only those throws that exceeded 50 metres. This resulted in the longest 150 throws being selected for the analysis.

The following parameters were obtained from the EVT analysis:

$$\beta = 15,$$

$$\gamma = 3.$$

(ii) Determine the percentage of javelin throws that would be expected to exceed 70 metres next year. [4]

(iii) Comment on the limitations of this analysis. [4]

Answer:

(i)

By fitting a distribution across the whole data range, the single distribution chosen may be a good overall fit of the data but could be a poor fit where there is little data, e.g. in the tails which are of primary concern.

EVT can be useful where we are particularly interested in the tail of a distribution and need to model that part accurately.

(ii)

The threshold is 50 metres.

P(X > 70)
= P(X > 70 given X > 50) * P(X > 50)

Now:
P(X > 70 given X > 50)
= P(Threshold exceedances > 20)

= 1 – P(Threshold exceedances ≤ 20)

= 1 – G(20), where $G(x)=1- (1 + \frac{x}{\gamma\beta})^{-\gamma}$ is the CDF for the Generalised Pareto Distribution

= 1 – [1 – $(1+\frac{20}{3*15})^{-3}$]

= 0.331816

So:

P(X > 70)
= P(X > 70 given X > 50) * P(X > 50)
= 0.331816 * (150/3000)

= 0.016591 = 1.6591%

(iii) There are a number of limitations with this analysis:

Not all throws are independent.
OR:
An example of a source of non-independence, e.g. each thrower will make multiple throws. [1]

Not all throws are identically distributed.
OR:
An example of a source of non-identical distribution, e.g. changing weather conditions, different abilities of throwers. [1]
There could be different throwers next year, compared to the year analysed. [1]
There could be trends in the distances thrown over the years (e.g. improvements in training techniques, improvements in javelin technology (e.g. lighter javelins)). [1]
Changes to rules and regulations might influence the distances thrown. [1]
Alternative thresholds should be analysed. [1]
The sample size is not particularly large. [1]
The generalized Pareto distribution is a limiting distribution and the actual distribution of the exceedances over any finite threshold will be different. [1]
[Marks available 8, maximum 4]

Q3.
Given Pareto distributions with parameters $\alpha = 2$ and $\alpha = 3$ (both with the same value of λ),
(a) Find the limiting density ratio.

(b) Which of the two has a thicker tail?

Answer:

(a) At the far end of the upper tail, the ratio of density functions:

$$\lim_{x\to\infty} \frac{f_{\alpha=2}(x)}{f_{\alpha=3}(x)} = \lim_{x\to\infty} \left\{ \frac{2\lambda^2}{(\lambda+x)^3} \Big/ \frac{3\lambda^3}{(\lambda+x)^4} \right\} = \frac{2}{3\lambda} \lim_{x\to\infty} (\lambda + x) = \infty$$

(b) The distribution with $\alpha = 2$ has a thicker tail.

Q4. R

Risk models: Parameter variability in a heterogeneous portfolio

Suppose that the Poisson parameters of 100 policies in a portfolio are not known but are equally likely to be 0.1 or 0.3 and claims are from a *Gamma*(750, 0.25) distribution.

It may be helpful to think of the above as a model of part of a motor insurance portfolio. It is supposed that some of the policyholders in this part of the portfolio are 'good' drivers and the remainder are 'bad' drivers. The individual claim amount distribution is the same for all drivers but 'good' drivers make fewer claims (0.1 pa on average) than 'bad' drivers (0.3 pa on average). It is assumed that it is known, possibly from national data, that a policyholder in this part of the portfolio is equally likely to be a 'good' driver or a 'bad' driver but that it cannot be known whether a particular policyholder is a 'good' driver or a 'bad' driver.

(i) Simulate 10,000 values for the aggregate claim amount from a policy chosen at random from the portfolio, and hence estimate the mean and standard deviation of the aggregate claims from a randomly chosen policy.

Hint:

Create a vector S of length 10,000. This vector will store the 10,000 simulations of aggregate claim amount from a policy chosen at random.

```
sims <- 10000
```

```
S <- numeric(sims)
```

Then simulate 10,000 values of the Poisson parameter, lambda by sampling from 0.1 and 0.3.

```
set.seed(123)
```

```
lambda <- sample(x=c(0.1,0.3), replace=TRUE, size = sims, prob
= c(0.5,0.5))
```

Then simulate 10,000 values for *N*, the number of claims for a policy chosen at random from the portfolio.

```
N <- rpois(sims,lambda)
```

Then for each of the 10,000 simulations, i.e. `for (i in 1:10000)`

Simulate $X_1,\ldots, X_N$ sand sum up $X_1,\ldots, X_N$ to arrive at aggregate claim amount.

```
for (i in 1:sims)
```

```
  {S[i] <-sum(rgamma(N[i], 750, 0.25))}
```

(ii) Comment on the mean and standard deviation you obtained in (i) compared with the theoretical mean and standard deviation of the aggregate claims from a randomly chosen policy.

Answer:

(i)

```
sims <- 10000

S <- numeric(sims)

set.seed(123)

lambda <- sample(x=c(0.1,0.3), replace=TRUE, size = sims, prob
= c(0.5,0.5))

N <- rpois(sims,lambda)

for (i in 1:sims)

  {S[i] <-sum(rgamma(N[i], 750, 0.25))}

> mean(S)

[1] 606.8383

> sqrt(var(S))

[1] 1373.769
```

(ii) Let S_i denote the aggregate claim for a randomly chosen policy.

$S_i|\lambda_i$ has a straightforward compound Poisson distribution with Poisson parameter λ_i

$E(\lambda_i) = 0.1*0.5 + 0.3*0.5 = 0.2$

$var\ (\lambda_i) = E(\lambda_i^2) - [E(\lambda_i)]^2 = 0.1^2*0.5 + 0.3^2*0.5 - 0.2^2 = 0.01$

Theoretical mean, $E(S_i)$

$= E(E(S_i|\lambda_i))$, by the law of total expectation

$= E(\lambda_i m_1)$, using formula for the mean of a compound Poisson distribution with Poisson parameter λ_i

$= E(\lambda_i)\, m_1$

$= 0.2*(750/0.25)$, note that $m_1 = E(X) = \alpha/\beta$ for $X \sim Gamma(\alpha, \beta)$

$= 600$

Theoretical variance, $var\,(S_i) = E\,(var\,(S_i\,|\lambda_i)) + var\,(E\,(S_i\,|\lambda_i))$, by the law of total variance

$= E\,(\lambda_i\, m_2) + var\,(\lambda_i\, m_1)$, using formulae for the mean and variance of a compound Poisson distribution with Poisson parameter λ_i

$= E\,(\lambda_i)\, m_2 + var\,(\lambda_i)\, m_1^2$

$= 0.2\, m_2 + 0.01\, m_1^2$

Now $m_2 = E(X^2) = var(X) + [E(X)]^2$

$= 750/(0.25^2) + (750/0.25)^2$, note that $var(X) = \alpha\,/\,(\beta^2)$ for $X \sim Gamma(\alpha, \beta)$

So $var\,(S_i) = 0.2 * 9{,}012{,}000 + 0.01 * (3000^2) = 1{,}892{,}400$

So theoretical standard deviation = sqrt(1,892,400) = 1,375.6

So for a randomly chosen policy, the mean and standard deviation of simulated aggregate claims (606.7 and 1373.8 respectively) are close to the theoretical mean and standard deviation (600 and 1,375.6). The difference is due to sampling error.