Modern Statistical Methods Assignment

James Aley

February 2016

Question 1

(a)

The Pareto distribution has probability density function $f(x)$ given by:

$$f(x) = \frac{\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}} \quad x \geq 0$$

To calculate the cumulative distribution function, we integrate this function over the range $[0, u]$:

$$\begin{aligned} F(u) &= \int_0^u f(x)\,\mathrm{d}x \\ &= \int_0^u \frac{\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}}\,\mathrm{d}x \\ &= \Big[-\beta^\alpha(x+\beta)^{-\alpha}\Big]_0^u \\ &= 1 - \frac{\beta^\alpha}{(u+\beta)^\alpha} \\ &= 1 - \left(\frac{\beta}{u+\beta}\right)^\alpha \end{aligned}$$

Now, to calculate the expected value, using integration by parts:

$$\begin{aligned}
\mathbf{E}\left[X\right] &= \int_0^\infty xf(x)\mathrm{d}x \\
&= \int_0^\infty \frac{x\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}}\,\mathrm{d}x \\
&= \left[-\beta^\alpha x(x+\beta)^{-\alpha} + \int \beta^\alpha (x+\beta)^{-\alpha}\right]_0^\infty \\
&= \left[-\beta^\alpha x(x+\beta)^{-\alpha} + \frac{\beta^\alpha}{1-\alpha}(x+\beta)^{1-\alpha}\right]_0^\infty
\end{aligned}$$

We can see by taking limits on the upper bounds, that we require $\alpha > 1$ in order for this integral to converge.

$$\lim_{x\to\infty}\left[-\beta^\alpha x(x+\beta)^{-\alpha} + \frac{\beta^\alpha}{1-\alpha}(x+\beta)^{1-\alpha}\right]_0^\infty = 0 \quad (\alpha > 1)$$

Therefore, substituting in $x = 0$ and subtracting from 0 gives us:

$$\begin{aligned}
\mathbf{E}\left[X\right] &= 0 - \frac{\beta^\alpha}{\beta^\alpha}\left(\frac{\beta}{1-\alpha}\right) \\
&= \frac{\beta}{\alpha - 1}
\end{aligned}$$

For the median, we need to find the value m that puts equal probability mass on either side of it. That is to say that $F(m) = \frac{1}{2}$

$$\begin{aligned}
F(m) &= \frac{1}{2} \\
1 - \left(\frac{\beta}{m+\beta}\right)^\alpha &= \frac{1}{2} \\
\frac{\beta}{m+\beta} &= \left(\frac{1}{2}\right)^{\frac{1}{\alpha}} \\
m &= \beta\left(\frac{1}{2^{\frac{1}{\alpha}}} - 1\right)
\end{aligned}$$

To calculate the variance, $\mathbf{Var}\left[X\right]$, we shall make use of the formula:

$$\mathbf{Var}\left[X\right] = \mathbf{E}\left[X^2\right] - \mathbf{E}\left[X\right]^2$$

We already have the first moment available from from calculating the mean earlier, so now we need to calculate the second moment, $\mathbf{E}\left[X^2\right]$ as follows. This time, we'll need to apply integration by parts twice successively.

$$
\begin{aligned}
\mathbf{E}\left[X^2\right] &= \int_0^\infty x^2 f(x)\,\mathrm{d}x \\
&= \int_0^\infty \frac{x^2\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}}\,\mathrm{d}x \\
&= \alpha\beta^\alpha\left[\frac{-x^2(x+\beta)^{-\alpha}}{\alpha} - \int \frac{-2x(x+\beta)^{-\alpha}}{\alpha}\right]_0^\infty \\
&= \beta^\alpha\left[-x^2(x+\beta)^{-\alpha} + \left(\frac{2x(x+\beta)^{1-\alpha}}{(1-\alpha)} - \int \frac{2(x+\beta)^{1-\alpha}}{(1-\alpha)}\right)\right]_0^\infty \\
&= \beta^\alpha\left[-x^2(x+\beta)^{-\alpha} + \frac{2x(x+\beta)^{1-\alpha}}{(1-\alpha)} + \frac{2(x+\beta)^{2-\alpha}}{(1-\alpha)(2-\alpha)}\right]_0^\infty
\end{aligned}
$$

This integral will only converge if we require that $\alpha > 2$, in which case we take limits for x as before:

$$
\lim_{x\to\infty}\beta^\alpha\left[-x^2(x+\beta)^{-\alpha} + \frac{2x(x+\beta)^{1-\alpha}}{(1-\alpha)} + \frac{2(x+\beta)^{2-\alpha}}{(1-\alpha)(2-\alpha)}\right] = 0
$$

Note that the requirement that $\alpha > 2$ is important both so that the x^2 component above grows more slowly than the denominator, and so that the final quotient is finite. Again, we substitute in $x = 0$ and subtract from the 0 obtained from this limit to get an expression for the second moment:

$$
\begin{aligned}
\mathbf{E}\left[X^2\right] &= 0 - \beta^\alpha\left(\frac{2\beta^{2-\alpha}}{(1-\alpha)(2-\alpha)}\right) \\
&= \frac{2\beta^2}{(\alpha-1)(\alpha-2)}
\end{aligned}
$$

Returning to the expression for $\mathbf{Var}\left[X\right]$, we have:

$$
\begin{aligned}
\mathbf{Var}\left[X\right] &= \mathbf{E}\left[X^2\right] - \mathbf{E}\left[X\right]^2 \\
&= \frac{2\beta^2}{(\alpha-1)(\alpha-2)} - \left(\frac{\beta}{\alpha-1}\right)^2 \\
&= \frac{2\beta^2\alpha(\alpha-1) - \beta^2(\alpha-2)}{(\alpha-1)^2(\alpha-2)} \\
&= \frac{\beta^2\alpha}{(\alpha-1)^2(\alpha-2)}
\end{aligned}
$$

(b)

To generate samples from the Pareto distribution, using uniform variables, we can invert the CDF obtained in the previous section as follows.

$$\begin{aligned} F(x) &= 1 - \left(\frac{\beta}{x+\beta}\right)^{\alpha} \\ \frac{\beta}{x+\beta} &= (1 - F(x))^{\frac{1}{\alpha}} \\ x &= \beta\left[(1 - F(x))^{-\frac{1}{\alpha}} - 1\right] \end{aligned}$$

This gives us an expression for $F^{-1}(x)$, the inverse cumulative distribution function:

$$F^{-1}(u) = \beta\left[(1 - u)^{-\frac{1}{\alpha}} - 1\right]$$

We may now use values for u, which should be realisations of $U \sim \text{Uniform}(0,1)$. These may be generated using a congruential generator, or any standard method for generating pseudo-random uniform numbers. The `runif()` function can be used to achieve this in R.

Note: R Plotting Requirements

The remainder of this document makes use of the `ggplot2` R package for rendering plots. The following snippet of code can be run to make this available in order to reproduce the plots:

```r
# Install ggplot if it's not already available
install.packages('ggplot2')

# Now load the functions from the ggplot2 package
require(ggplot2)
```

(c)

The following code snippet includes implementations of the Pareto distribution PDF, CDF and inverse CDF, as calculated in the previous section. Following those function definitions, we make use of them to simulate values from the Pareto distribution and visualise that simulation using the method outlined previously.

```r
#' Probability density function for the Paretro distribution, with
#' parameters alpha and beta
pareto.pdf <- function(alpha, beta, x) {
  (alpha * beta^alpha) / (x + beta)^(alpha + 1)
}
```

```r
#' Cumulative distribution funtion for Paretro distribution with paremeters
#' alpha abd beta.
pareto.cdf <- function(alpha, beta, u) {
  1 - (beta / (u + beta))^alpha
}

#' Inverse cumulative distribution function for the Pareto
#' distribution. Alpha and Beta are distribution parameters,
#' q should be a value between 0 and 1.
pareto.inv.cdf <- function(alpha, beta, q) {
  beta * ((1 - q)^(-1/alpha) - 1)
}

## Compare simulated values to the density function
pareto <- data.frame(simulated = pareto.inv.cdf(3, 100000, runif(10000)))
ggplot(pareto, aes(simulated)) +
  geom_density(colour="red", fill="red", alpha=0.1) +
  stat_function(fun = function(x) {pareto.pdf(3, 100000, x)},
                colour="blue", geom="area", alpha=0.1, fill="blue") +
  xlab("X") +
  ylab("Density") +
  xlim(0, 100000)
```

Note that in Figure 1 we're plotting the real Pareto PDF against the *density* of the simulated values, so that they are on comparable scales. We're also focusing on the range for X in 0 to 100,000, before entering the very long tail. We can see that the simulation fits the PDF reasonably well, and is mostly unable to reach the very low extremes of the real PDF.

(d)

Insurance claims are likely to me well-modeled by the Pareto distribution, as it is very heavy-tailed. It is able to realistically represent the very low probability of large claims, corresponding to extreme loss for the insurer. Insurance claims are also likely to follow the *Pareto Principle*, in that most of the value of the payments made will be to relatively few (large) claims.

Question 2

The following R code snippet simulates possible outcomes for 1000 customers, each with a claim probability of 0.1 in a year. We assume the claims X have distriubtion $X \sim \text{Pareto}(3, 100000)$ as before.

We also note, given the restriction that each customer can only make a single claim in a year, we essentially have a Bernoulli outcome variable for whether

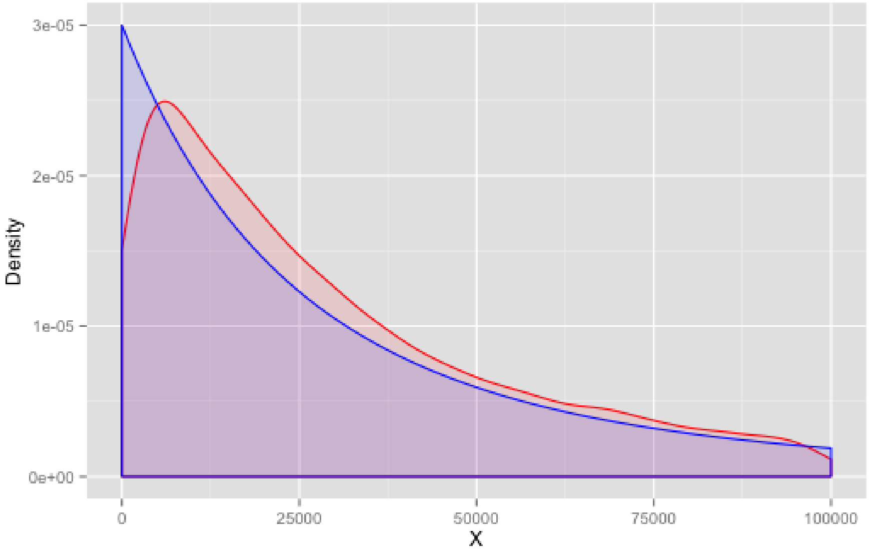

Figure 1: Simulated value density (red) plotted against exact Pareto PDF (blue).

or not each customer makes a claim with parameter $p = 0.1$. Modeling the event of a claim occurring as 1 and not occurring as 0, we can extend this to n customers be using a Binomial distribution. If we let N represent the number of claims, then:

$$N \sim \text{Bin}(1000, 0.1)$$

The code below uses this to model the number of claims, `n_claims` in each simulation iteration. That variable tells us the number of uniform variables to draw and pass on to our inverse Pareto CDF derived previously to model the value of claims for one year.

```
#' Returns a vector of length n_runs containing simulated values
#' for the year end assets.
simulate.year.end.assets <- function(n_runs = 10000,
                                     start_assets = 250000,
                                     n_customers = 1000,
                                     premium = 6000,
                                     claim_prob = 0.1) {
  # Accumulate year-end assets in a vector
  outcomes <- numeric(length = n_runs)
  for(i in 1:n_runs) {
    # Number of claims this year, in this simulation
    n_claims <- rbinom(1, n_customers, claim_prob)

    # Calculate year-end profit/loss for this number of claims, by
    # drawing from Pareto distribution
    outcomes[i] <- start_assets +
      (n_customers * premium) -
      sum(pareto.inv.cdf(3, 100000, runif(n_claims)))
  }
  return(outcomes)
}

# Reformat simulation data into a data frame for plotting
# a histogram of the simulation outcomes
bankruptcy <- data.frame(assets=simulate.year.end.assets())
ggplot(bankruptcy, aes(assets)) +
  geom_density(colour="blue", fill="blue", alpha=0.1) +
  geom_vline(xintercept = 0, linetype="longdash") +
  xlab("Assets at Year End") +
  ylab("Probability Density")

# Calculate the probability of bankruptcy, i.e. the proportion
# of outcomes where year end profits were less than zero.
sum(bankruptcy[’assets’] < 0) / n_runs
```

In one simulation, we see a probability of bankruptcy probability of 0.0954, from the outcome distribution shown in Figure 2.

Question 3

(a)

The following code snippet makes use of the simulation function we defined in the previous section, calling it with a different premium price parameter many times to produce the plot in Figure 3.

```
#’ Esimate the probability of bankruptcy for a premium of given price
bankruptcy.prob <- function(premium) {
  N <- 10000
  assets <- simulate.year.end.assets(premium=premium, n_runs=N)
  sum(assets < 0) / N
}

premiums <- seq(from=5500, to=8000, by=250)
premium.data <- data.frame(premium.price = premiums,
                           bankruptcy.prob = sapply(premiums, bankruptcy.prob))
```

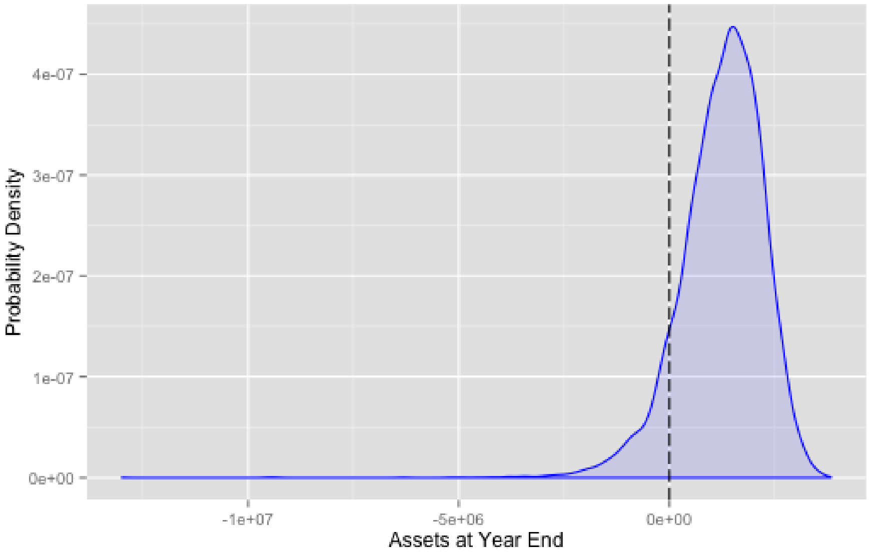

Figure 2: Simulated year-end profit or loss.

```
ggplot(premium.data, aes(x=premium.price, y=bankruptcy.prob)) +
  geom_line(colour="blue") +
  stat_hline(yintercept = 0.02, linetype="longdash") +
  xlab("Premium Price (GBP)") +
  ylab("Probability of Bankruptcy")
```

As indicated by the dashed line in Figure 3, we should set the premium to £7000 or higher in order to see a probability of bankruptcy below 2%.

(b)

Again, making use of the simulation functions defined previously, the following code snippet reruns the simulation for varying claim probabilities to produce the plot shown in Figure 4.

```
bankruptcy.prob.claim <- function(claim.prob) {
  N <- 10000
  assets <- simulate.year.end.assets(claim_prob = claim.prob, n_runs=N)
  sum(assets < 0) / N
}

claims <- seq(from=0.05, to=0.15, by=0.005)
claim.data <- data.frame(claim.prob = claims,
```

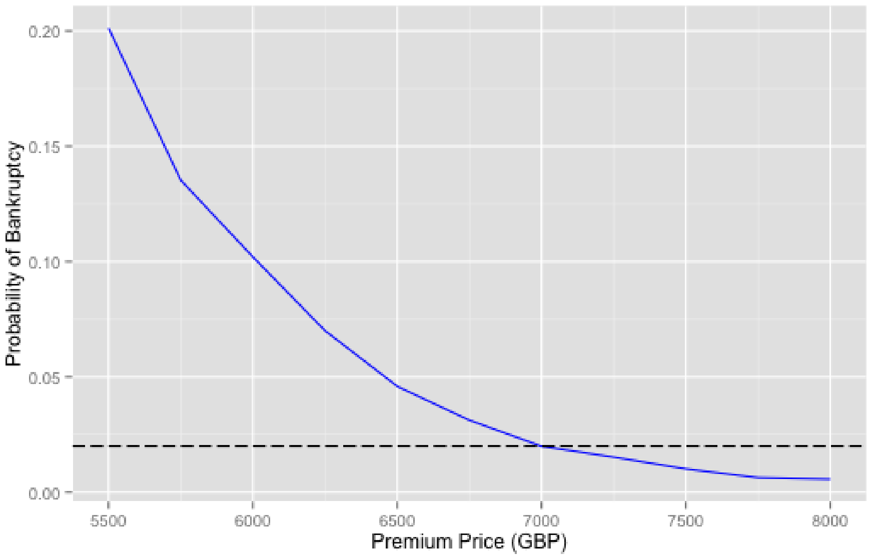

Figure 3: Probability of bankruptcy for changing premium prices.

```
                        bankruptcy.prob =
                          sapply(claims, bankruptcy.prob.claim))

ggplot(claim.data, aes(x=claim.prob, y=bankruptcy.prob)) +
  geom_line(colour="blue") +
  stat_hline(yintercept = 0.02, linetype="longdash") +
  xlab("Claim Probability") +
  ylab("Probability of Bankruptcy")
```

Holding the premium price fixed at £6000, as per the original specification, we can see from Figure 4 that the company can only expect a bankruptcy probability below 2% if the claim probability is below 0.08.

Question 4

This report summarises an investigation into company year-end asset values and provides recommendations accordingly. The investigation in question involved running simulations based on information we currently hold on customers and insurance claim behaviour. The following assumptions have been assumed to be accurate throughout simulations:

- Number of customers: 1,000
- Annual premium charged: £6,000

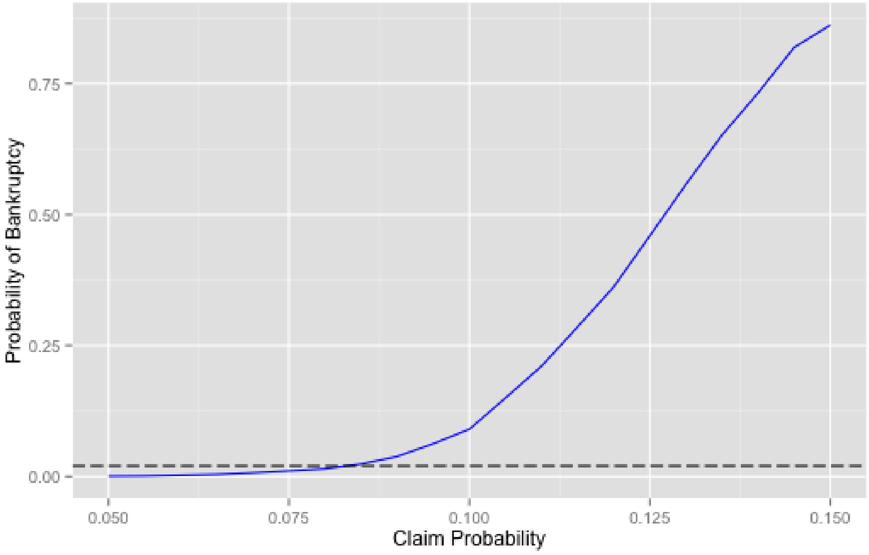

Figure 4: Bankruptcy probability for varying customer claim probabilities.

- Current asset value: £250,000
- Customers can only make one claim per year
- 10% probability of a customer making a claim
- Claims can be reasonably modeled by a Pareto distribution with parameters $\alpha = 3$ and $\beta = 100,000$

Based on these assumptions, the mean asset valuation at the end of the year in simulations was approximately £1250000. However, variability was seen to be very high, and in fact a 95% confidence interval for asset valuation was (-£900000, £2900000). That is to say that though substantial profits are entirely likely, we cannot rule out the possibility of a significant loss or bankruptcy at the 95% level of significance.

In simulations, we found that there is currently about a 10% chance of bankruptcy with current pricing and claim characteristics. This is worryingly high, and certainly something that the company should look to address in order to secure sustainable business. Various factors have been investigated in simulations to assess impact on these metrics and are summarised in following sections.

Premium Price

Unsurprisingly, increasing the customer premium price can be shown to reduce probability of bankruptcy and increase predicted profits according to simulations. Figure 3 shows how increases in the current premium price reduce the probability of bankruptcy. If we work with a target of reducing the probability to 2%, indicated by the dotted line, then increasing the premium price to £7000 would be sufficient.

The corresponding mean profit simulated was £2250000 with 95% confidence limits of (£166000, £3900000). It should of course be noted however that this change would inevitably cause some customers to revisit their choice of insurance supplier, and we would inevitably lose some customers. This impact of the number of customers buying insurance is investigated in a subsequent section.

Claim Probability

The probability of customers making a claim has a clear impact on profits and risk of bankruptcy too, as shown in Figure 4. Reducing the probability of a claim from the current 10% to 8% would also reach the 2% bankruptcy risk target. Impact on profit estimates are comparable to the suggest increase in premium price from the previous section.

It should be noted, however, that this factor is likely more difficult for the company to control. Various measures could be taken to limit what customers are able to make claims for, but again this is likely to cause customers to choose other suppliers in some circumstances.

Number of Customers

Perhaps the least risky option in terms of unpredictable and potentially undesirable consequences would be to increase the number of customers buying insurance. Figure 5 illustrates the impact of this initiative on bankruptcy risk via simulations again.

The 2% target can be reach by increasing our customer base four times to 4,000. If it is possible to achieve this through advertising initiatives then we could substantially improve profits and decrease bankruptcy risk without putting our existing customer relationships at risk.

Recommendations

Any of the above factor increments could improve on the current situation, but some difficult judgment calls need to be made on the trade-offs we expect to see. A few key pieces of information that would help in this respect are:

- How many customers do we expect will cancel their policy per unit increase in premium price?

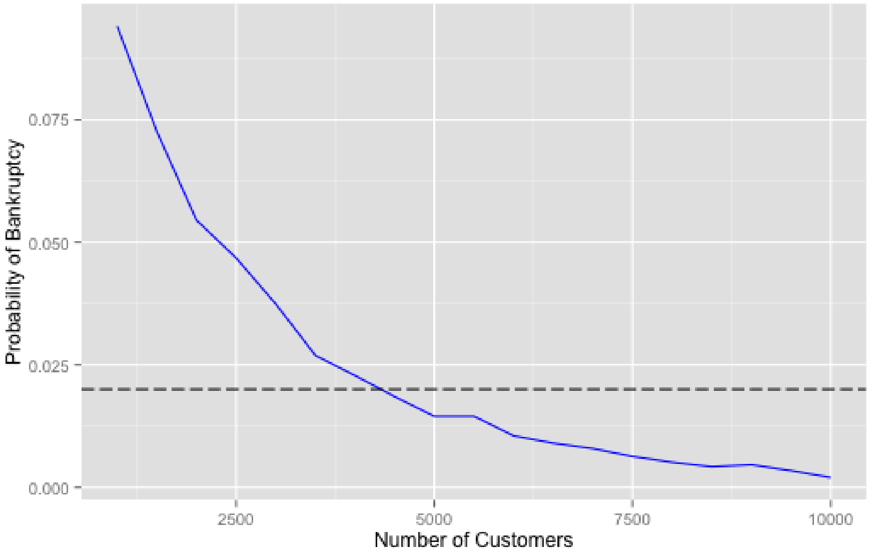

Figure 5: Bankruptcy probability for increasing customer numbers

- Are there are any policy changes that can be made to reduce the probability of a claim from each customer, and again, what would the impact be in terms of cancellations?
- What would the cost be of advertising initiatives required to increase customer base to the suggested levels?

With this information, further modeling could be carried out to provide more reliable suggestions. However, in the absence of this research it seems that it might be a good idea to hedge bets by instead designing further insurance products (assuming they can reach the market in a reasonable time-frame) that exhibit one of the following:

- Attract much higher customer volumes with lower claim probabilities, likely at the cost of a lower premium price. That is, an "economy" or "entry-level" policy for a new class of customer. Equivalently, we could design this policy to have a cap on the maximum payout or a more predictable distribution of claims to limit losses.
- A "high-end" option with a much higher premium, again targeting new customers. We expect claims with higher probability here and offer a high cap on maximum claim, but cover this with a bigger premium.