Required Sample Size in Capital Modeling

Kevin Zhang, PhD, FCAS

The Cincinnati Insurance Company

Contents

- Description of a capital model
- Why is the sample size important?
- Mathematical preliminary: limit distribution of a sample statistic
- A realistic example: required sample size for estimating the following statistics
    - The mean
    - Value-at-Risk (VaR)
    - Tail-Value-at-Risk (TVaR)
    - Multiple measures
- Tail behavior of modeled surplus is dominated by catastrophe losses
- Unsolved problems

A capital model is used to determine the required capital

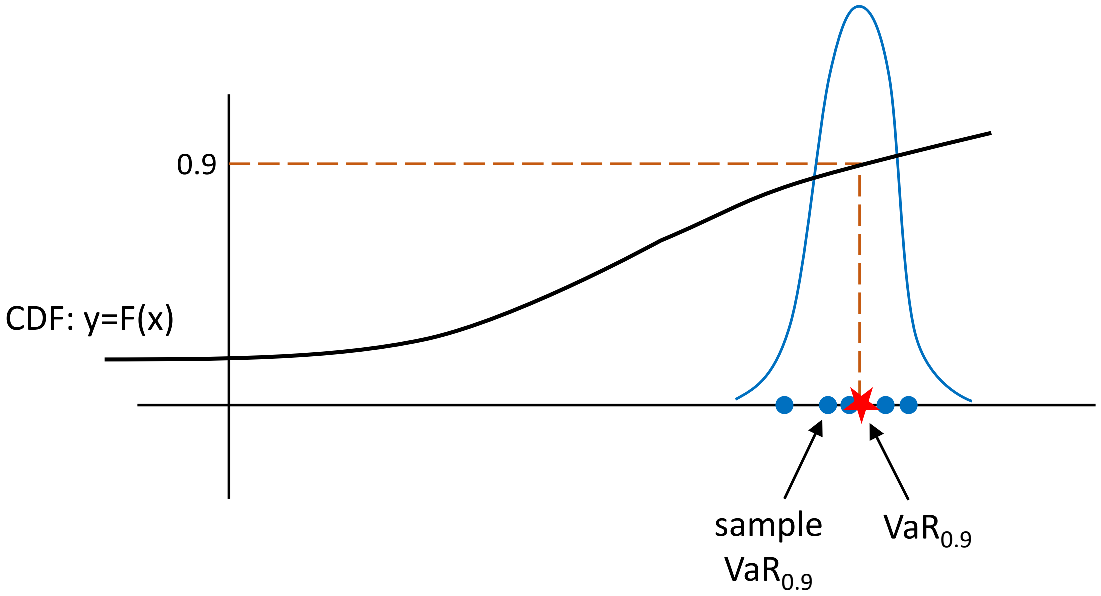
- A one-year stochastic model: we model the yearend surplus change, and compute the required capital as a VaR or TVaR of the surplus change distribution
- Investments and catastrophes contribute the greatest risk
  - Other risks are less volatile in one year
- To estimate VaR (TVaR), we run a simulation to get a sample, then find the corresponding sample VaR (TVaR)
- Question: How many paths should we simulate so that a sample VaR (TVaR) is a good estimate of the true VaR (TVaR)?
- Note: in this study our question is not whether a model is a good model, but whether a sample is a good sample

Sample size is a big problem in capital modeling

- Capital concerns the deep tail, e.g., 1-in-1000-year surplus loss
- A distribution of cat losses has a fat tail
- High degree of precision is needed, since risk metrics are reported to the BoD, rating agencies and regulators and are used in making business decisions
- In other fields involving random sampling, a sample size of 10,000 is considered large. In capital modeling, 100,000 is often not enough

To estimate a risk metric, we run a simulation and calculate the corresponding metric of the sample

- Intuition: When a sample is large, the empirical distribution of the sample is near the true distribution, and any sample statistic is near the true statistic
- Mathematics: Sample statistics, with varying seeds, form a normal distribution with a small variance and centered at the true statistic. Increasing the sample size will decrease the variance



We are all familiar with how sample means are distributed around the true mean

- Central Limit Theorem: Draw a random sample of size n from a distribution with mean μ and variance σ^2. If n is large, then the sample mean $\bar{X}$ is (approximately) normally distributed with mean μ and variance $\frac{\sigma^2}{n}$
- If we desire $\bar{X}$ to fall in the interval $(\mu - a, \mu + a)$ with 95% probability, then

$$n \geq \left(\frac{1.96\sigma}{a}\right)^2$$

Go to slide 12

There is a similar formula for VaR

- To estimate $\mathrm{VaR}_p(X) \equiv x_p$, we draw a sample of size n and find the $[n(1-p)]^{\text{th}}$ largest value $\hat{x}_p$
- If we desire $\hat{x}_p$ to fall in the interval $(x_p - a, x_p + a)$ with 95% probability, then

$$n \geq \left(\frac{1.96}{af(x_p)}\right)^2 p(1-p)$$

where $f(x)$ is the probability density function of X

Go to slide 15

… and for TVaR

- To estimate $\mathrm{TVaR}_p(X) \equiv t_p$, we draw a sample of size n and calculate the sample TVaR

$$\hat{t}_p = \frac{1}{n(1-p)} \sum_{x_i > \hat{x}_p} x_i$$

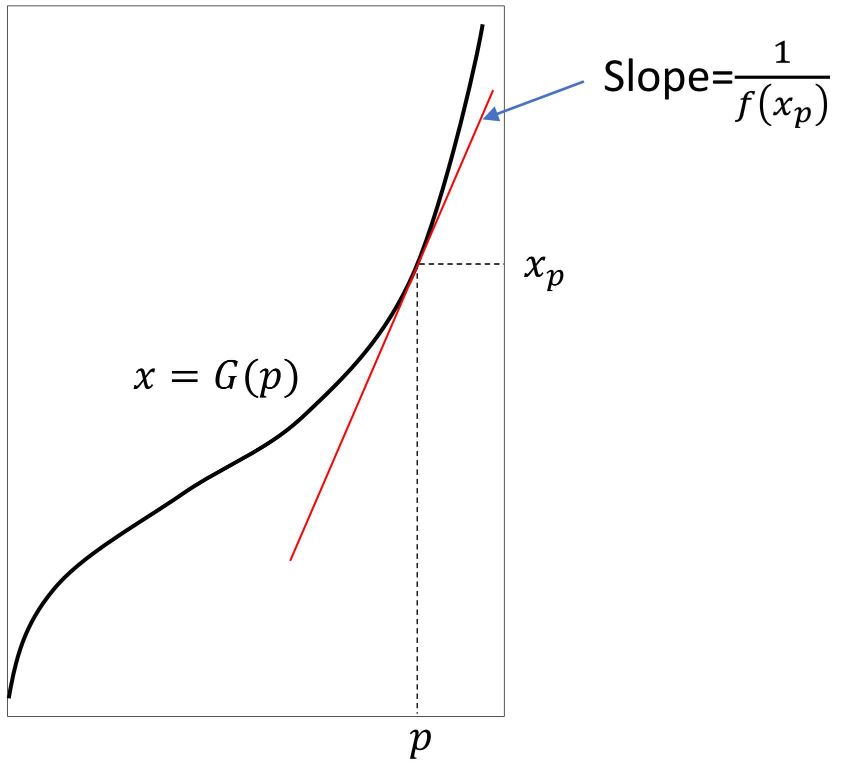
- If we desire $\hat{t}_p$ to fall in the interval $(t_p - a, t_p + a)$ with 95% probability, then

$$n \geq \left(\frac{1.96}{a(1-p)}\right)^2 V\left(\left(X - x_p\right)^+\right)$$

where V means the variance

The quantile function is useful for understanding sample-size formulas related to VaR and TVaR

- The quantile function for a CDF $y = F(x)$ is $x = F^{-1}(p) \equiv G(p)$
- To use the formula in slide 7 we need to estimate $\frac{1}{f(x_p)}$, which is the slope $G'(p)$



How to estimate the unknown parameters σ, $f(x_p), V\left((X - x_p)^+\right)$?

- Estimate parameters from an arbitrary sample
  - Okay to run fewer paths, since only need rough estimates
  - Slightly conservative estimates are desirable
- May use prior knowledge about the distribution
  - E.g., if a stock portfolio is 5% larger than previous year, then its σ is likely also 5% larger
- May simulate a component instead of the full model
  - E.g., surplus loss is dominated by cat loss at the tail, which means the CDF curves of surplus and cat have similar shape. Then $f(x_p)$ or $V\left((X - x_p)^+\right)$ of the former can be estimated with that of the latter (see slide 20)

A realistic example

- A midsized P&C company
  - o $1B property premium; $160M mean cat loss
  - o $5B stock & bond portfolio; 4% mean investment gain
- Modeled surplus loss SL is a sum of
  - o Modeled cat loss CL
  - o Modeled investment loss IL (negative investment gain)
  - o Other items including reserve change, premium, expense and non-cat loss, which altogether vary modestly in one year
- It suffices to calculate required sample size for the “reduced” surplus loss

$$SL = CL + IL$$

Purpose of solving the example

To demonstrate:

- What questions may be asked?
- What questions may be answered?
- What results may be expected?
- What problems are yet to be solved?

How big a sample is required for estimating the mean?

1. Run a simulation and compute the sample standard deviation (in $million)

$$\hat{\sigma}_{CL} = 97.6 \qquad \hat{\sigma}_{IL} = 278.9$$

2. Set error tolerance to be about 1% of mean

$$a_{CL} = 2 \qquad a_{IL} = 2$$

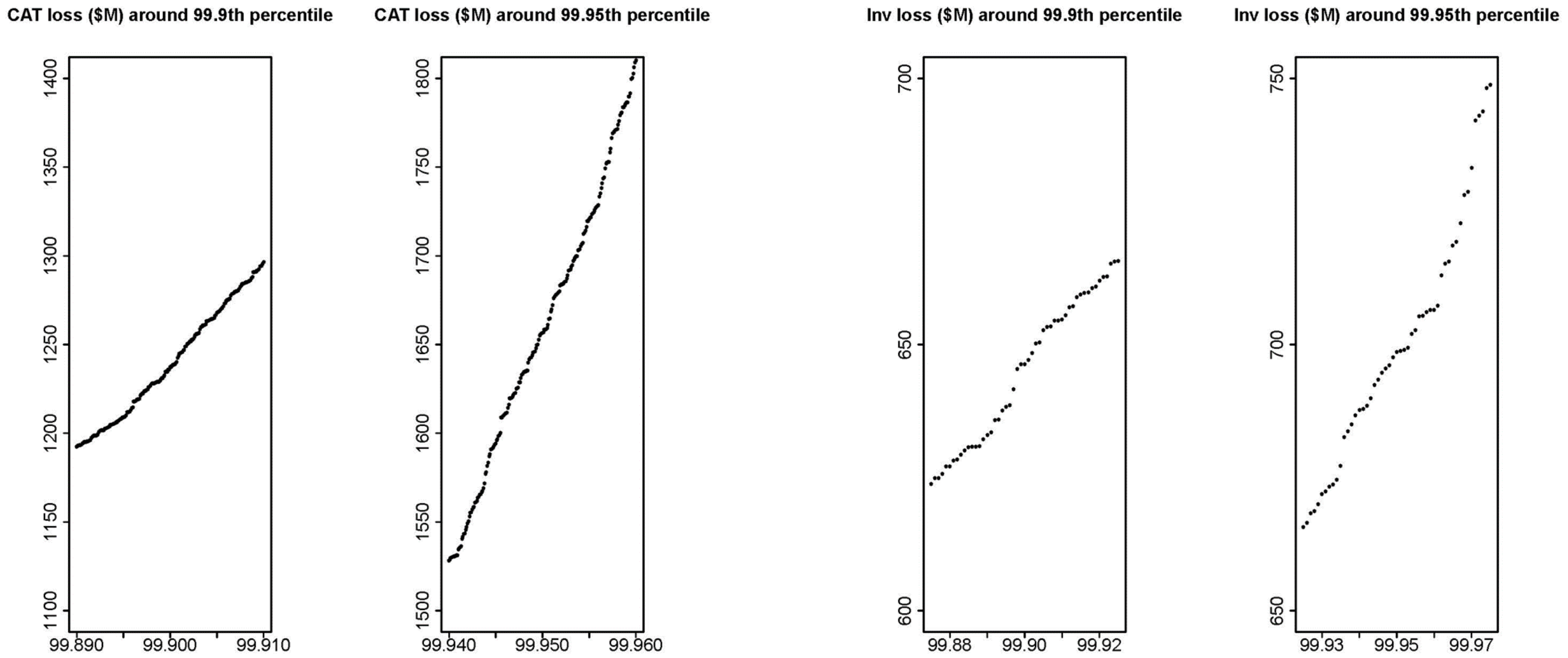
3. Use the formula in slide 6 to compute the required sample size

$$N_{CL} = 9{,}100 \qquad N_{IL} = 74{,}700$$

4. Some conclusions
    - The investment loss has a larger SD than the cat loss, although the latter has greater tail risk
    - It requires more paths to estimate the mean for the investment loss than for the cat loss

We need the slope of $G(p)$ to calculate sample sizes for VaR. It is easy to visualize slopes on a sample plot

1. Run an arbitrary simulation to estimate slopes @ 99.9% and 99.95%



How big a sample is required for estimating VaR?

2. Select an error tolerance $a = 100$ for both CL and IL at both 99.9% and 99.95%
3. Use the formula in slide 7 to compute the required sample size

$$N_{CL}(99.9\%) = 149{,}700 \quad N_{IL}(99.9\%) = 3{,}200$$
$$N_{CL}(99.95\%) = 309{,}800 \quad N_{IL}(99.95\%) = 4{,}300$$

4. Some conclusions
   - The cat loss has much greater slopes (on the quantile function) than the investment loss
   - It requires many more paths to estimate VaRs for the cat loss than for the investment loss
   - Sample size of 100,000 is not enough for accurately estimating cat VaRs, but more than enough for investment VaRs

What is the error range for a 100k-path simulation

- Given sample size 100,000 we can compute error size a (with 95% confidence) using the same formula

$$a_{CL}(99.9\%) = 122 \quad a_{IL}(99.9\%) = 18$$
$$a_{CL}(99.95\%) = 176 \quad a_{IL}(99.95\%) = 21$$

- For the cat loss, the error ranges are about 10% of the VaRs – not acceptable

We usually use TVaR at lower return periods than we do VaR. We will estimate TVaR at 99.8% and 99.9%

1. Run a simulations of cat and investment and compute from the samples the variance $V\left(\left(X - \mathrm{VaR}_p(X)\right)^+\right)$ at $p = 99.8\%$ and $p = 99.9\%$
2. Select error bound $a = 100$
3. Use the formula in slide 8 to compute the required number of paths

$$N_{CL}(99.8\%) = 172{,}300 \qquad N_{IL}(99.8\%) = 2{,}500$$
$$N_{CL}(99.9\%) = 456{,}600 \qquad N_{IL}(99.9\%) = 4{,}200$$

   At the same point (99.9%), required sample size is larger for TVaR than for VaR

4. If we only run 100,000 paths, error sizes (at 95% confidence) are

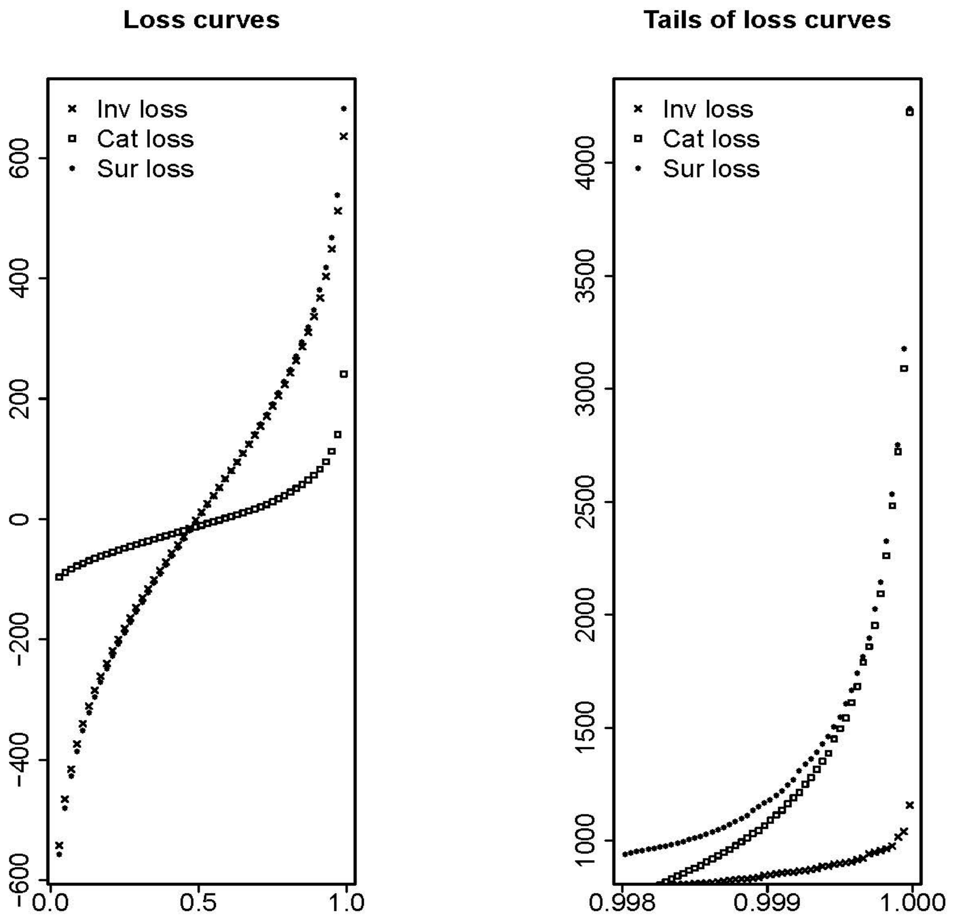
$$a_{CL}(99.8\%) = 131 \qquad a_{IL}(99.8\%) = 16$$
$$a_{CL}(99.9\%) = 214 \qquad a_{IL}(99.9\%) = 20$$

What other questions can we answer?

- What is the required sample size for simultaneously estimating two or more metrics?
  - Means of CL and IL
  - VaRs of CL at 99.5%, 99.9%, 99.95%
  - VaR and TVaR of CL at 99.9%
- Mathematically, these are derived by using a multivariate normal distribution to approximate a joint distribution of several samples

Among the components of the surplus loss, cats dominate the tail behavior, while investments determine the overall variability

We graph shifted cat loss CL-E(CL), investment loss IL-E(IL) and surplus loss SL-E(SL) in the same window



The relationship helps determine a sample size for estimating VaRs and TVaRs of the surplus loss

- To estimate the mean of SL
  - From $\hat{\sigma}_{CL} = 97.6$ and $\hat{\sigma}_{IL} = 278.9$ we get $\hat{\sigma}_{SL} = 295$, assuming CL and IL are independent
  - $\hat{\sigma}_{SL}$ is slightly larger than $\hat{\sigma}_{IL}$, so the sample size required for approximating E(SL) is slightly larger than that for approximating E(IL)
- To estimate VaR and TVaR of SL
  - A slope on the CL curve is larger than that on the SL curve at the same p. So, a sample size large enough for $\mathrm{VaR}_p(CL)$ is also large enough for $\mathrm{VaR}_p(SL)$
  - The same can be said for TVaR

Questions for future research

- What to do if the sample size is limited by hardware or software constraints?
  - If we are allowed to choose seeds, then we may run many simulations with different seeds, and average out the sample statistic from all simulations
- What is a required sample size for accurately estimating co-TVaRs?
  - Very important. But I have not seen any research

Appendix: terminologies regrading sampling

- Drawing a sample with n sample points $\Longleftrightarrow$ running a simulation with n paths
  - n is called the sample size
- Drawing k independent samples (each of size n) $\Longleftrightarrow$ running k simulations (each of n paths) with independent seeds
  - $k \times n$ sample points (or paths) in total