An Actuarial Perspective on Cyber Risk Management

I recently passed the CompTIA Security+ exam and have continued my education by reading many Cyber Security related books. Two of my most recent ones are Measuring and Managing Information Risk: A FAIR Approach (Jack Freund and Jack Jones) and How to Measure Anything in Cybersecurity Risk (Douglas Hubbard and Richard Seiersen). Having spent 18 years performing Property & Casualty Actuarial work, I was very familiar with many of the techniques discussed by the authors.

My analysis in this paper is intended as an overview of my findings of these books and their overlap with the actuarial field. One objective is to provide readers with a new way to look at loss data. In general, there is no singular way to approach data, though there is usually some consensus of generally accepted practices. Later in this paper, I will discuss an actuarial term called trend; in the field of actuarial science, some hold the opinion that there is no such thing; in fact, when I was taking my exams, this alternate view was one of the required readings. Similarly in Cyber Risk Management, some say you can't quantify the risk. One can only show their work, believe in their work and be prepared to defend their work. Since the models used are not as simple as 1+1=2 and are usually based on some amount of assumptions, a solid argument for those assumptions is worth its weight in gold.

Freund and Jones also point this out where they discuss the lack of standard definitions. For example

- Risk defined by Freund and Jones is *"The probable frequency and probable magnitude of future loss"*
- Risk in my insurance risk management courses is broken up into Pure and Speculative Risk. Pure Risk has either a financial loss or it doesn't. Speculative adds a third, such as an investment gain (or loss).

In the following table, I mapped the traditional risk management loss categories to those listed in the FAIR taxonomy. Overall they are very similar, just a bit more granular and with different names.

FAIR RM Category	Traditional RM Category
Productivity	Casualty
Response	Casualty
Replacement	Technological
Competitive Advantage	Market
Fines & Judgement	Political
Reputation	Market

What first spurred me to write this paper is the familiarity of the formulas used in the FAIR formula. It is very similar to the pricing formulas I used while working as a Pricing Actuary. For example, the FAIR formula (definition of risk) is:

LEF x LM = Risk

LEF = Loss Event Frequency

LM = Loss Magnitude

From my insurance point of view, the formula is similar to that for the “Pure Premium” or “Loss Cost”. This is the part of the insurance rate that is exclusively made up of the expected losses prior to any expenses and profit being added to it. Since we are only looking for the expected losses for a given IT risk, no discussion for converting the Loss Cost to a full premium rate is necessary for this discussion. Though in future discussions there might be some need for it; for example, income tax effects of tax-deductible losses. The following is the formula for pure premium:

Freq x Sev = PP

Freq = Frequency of losses

Sev = Severity of losses

PP = Pure Premium

Now we will deconstruct this further and use a definition for each.

Freq = # Claims / # Exposures (how many claims per exposure)

Sev = $ of Losses / # of Claims (how much each claim costs on average)

(A quick definition of exposure is a unit of measure for a given risk. In automobile insurance it is one car year)

Now if we multiply out:

(# Claims/ # Exposures) x ($ of Losses / # of Claims) = PP

Now we reduce the calculation to:

$ of Losses / # Exposures = PP

So in simple terms, the average loss component for a given exposure is the average number of losses per exposure. For example, if we have $100 in losses for 1,000 exposures, the PP would be $0.10, so each insured would pay 10 cents per exposure (plus an expense & profit provision),

While you can just take the final calculation of "$ of Losses / # Exposures" and call it a day; your model will be better off by using the deconstructed version to provide a better degree of control over the final calculation.

In its most basic form, an actuary would calculate the value for each of the following three variables:

# Claims

# Exposures

$ of Losses

In general, these are "known" quantities; however, this is where experience and judgment come into play. When I say known, if you look at any point in time, you know the quantities as of the last reporting date. What about the claims or losses that have been "Incurred but not Reported", IBNR, or where just a $1 reserve is set up for the claim? Just like in Cyber Security where a threat actor has already infected your system, but you haven't found it yet, insureds usually have 2 years to file a lawsuit. Or, the initial estimate for a loss could be way below the final outcome.

Generally, the number of exposures is fairly steady and not too much adjustment is necessary with this variable. The # of Claims could have some positive adjustment. The $ of losses is where the majority of adjustments will be made and will, therefore, have the largest impact on final results.

I've always thought that you can give 100 actuaries the same data and can come up with 100 different recommendations. The majority would hover around a central tendency and if plotted would form a typical bell curve with a small standard deviation. I reason that this is a small scale version of a Monte Carlo simulation. Each actuary is free to use their judgment in deciding what values to use when given a range of values and need to determine their estimate; i.e. do they choose the mean, median, mode, weighted average, the list goes on. The actuary can back up their decision and hence the estimate is valid.

When looking at data, the data needs to be standardized and of the same time period(s). For example, if looking at policies with a deductible, the amount of the deductible must be accounted for, otherwise, incorrect assumptions will be made.

There are many factors used to develop final losses. One factor is for the addition of "trend", think of this as inflation, but it is technically not. In Table 1, look at the amount the company paid in losses after a $1,000 deductible. The paid losses appear to be increasing at a fairly good clip; though the rate of change is decreasing.

Table 1

Year	Actual Paid Loss	Paid Change
1	$ 1	
2	$ 101	10000%
3	$ 211	109%
4	$ 332	57%
5	$ 465	40%

Now, in Table2, losses are adjusted for the $1,000 deductible.

Table 2

Year	Total Adj. Loss	Actual Change
1	$ 1,001	
2	$ 1,101	10%
3	$ 1,211	10%
4	$ 1,332	10%
5	$ 1,465	10%

Notice now, the change is 10% year after year. Had no adjustment been made, the final estimate for losses could be way off depending on the value selected for the trend if based off Table 1.

Say, based on your judgment of the trend in Table 1, you believe the growth in year 6 will be 25% which would give a year 6 loss of $1,581 ($465 x 1.25 + 1000), when in actuality it would be $1,612 (1,465 x 1.1).

Another relevant insurance term is “Tail”. Tails are usually considered long or short. In a short-tailed risk, such as automobile collision insurance, most of the losses are known very quickly. People want their money fast for the damage to their car, so they file a claim quickly, and since a car is a commodity, the value is known, and the claim is paid quickly. A long-tailed risk is usually a liability type risk, where even though the claim might be filed quickly, it could take years for the final outcome. Actuaries have tools to estimate the development over time.

Similarly in Cyber Risk Management, there are short and long-tailed risks. One big difference between Insurance and Cyber is typically the amount of information on how the losses develop over time. Individual insurance companies have their own loss data, however, many pool their data together to exploit the law of large numbers because their individual data can be statistically insignificant. NCCI, for worker’s compensation, is one such company, where member companies send their data and use the pooled results to adjust their own rates.

I bring these up because if you are looking at several years of data and trying to project what they might be in the future, each prior year must be "fully developed" to assume they were incurred in the time being analyzed. So, are your costs per incident increasing over time and is your business growing? Both of these would need to be modeled to provide a more accurate picture.

So, how can a company, especially a small one rely on their limited (if any) data since no real pooling mechanism exists for Cybersecurity. Granted there are some incident databases, that could provide some guidance. These databases to my knowledge are fairly limited in scope since they are either based on publicly reported data or voluntary in nature. Whereas in insurance, all the data for NCCI member companies are pooled and thus is very credible, or statistically significant. Some companies choose not to participate in such pooling mechanisms due to their sheer size and to provide a competitive advantage.

The following tables show the development of the losses from Target and TJ Maxx found in their 10K reports (annual reports filed with the SEC). As expected both companies highlighted these losses in different ways. Target called the intrusion "Data Breach" and provided insurance recoveries. TJ Maxx called the breach "Computer Intrusion and provided corporate reserves.

Target 2013 Breach 10K Reported Losses ($ millions)

	2013	2014	2015	2016	2017
Cumulative Losses Paid	$61	$252	$291	$291	$291
Insurance Recoveries	$44	$90	$90	$90	$95
Losses after Insurance	$17	$162	$201	$201	$196

Data from 10-K, Notes to Consolidated Financial Statements

TJ Maxx 2007 Breach 10K Reported Losses ($ thousands)

	2007	2008	2009	2010	2011	2012	2013	2014	2015
Incurred Losses to Date (1)	$4,960	$164,160	$151,160	$151,160	$149,610	$149,610	$149,610	$149,610	$149,610
Provision for Computer Intrusion Costs (2)	$4,960	$201,982	$171,482	$171,482	$159,932	$159,932	$159,932	Not Reported	

1) Data from 10-K, Schedule II—Valuation and Qualifying Accounts

2) Data from 10-K, Consolidated Statements of Income

Several key elements I believe are noteworthy.

- ✓ Both have very large developments from the 1st report in a 10K to the next 10K. I believe the timing of the report does impact this as the companies were coming to grips with the scope of the losses.
- ✓ Not much development happens after the third report. In fact, both companies have negative growth in the latter years:
    - Target has a $5 million insurance recovery in 2017; and,
    - TJ Maxx lowered reserves in 2011.

To an actuary, credibility is how much faith you have in your data; it is a formula based number between 0 and 1 and is typically based on the number of claims a risk has. A typical formula used in insurance rate making is:

Company# x Credibility + Pooled# x (1-Credibility) = Final#

Where:

Company# is the company's experience

Pooled# is the pooled industry's experience

With this formula, a company weighs its own experience with the industry's experience to formulate a weighted average. If a company has 100% credibility in its experience then Company# = Final#, otherwise it is a weighted average of the two.

This is why I believe that the FAIR model uses 90% Confidence Intervals; because most companies' loss history would be statistically insignificant, yet the potential for large losses does exist. Or vice versa, they could have had a large loss, put appropriate controls in place to limit future losses thereby changing their overall exposure.

To put it another way, I believe that the credibility in actuarial terms is related to the overall size of the confidence interval (CI).

High credibility = Low variation (tall and steep distribution) = Narrow CI

Low credibility = High variation (low and wide distribution) = Wide CI

One big difference I've noticed between the two risk management realms is how certain expenses are accounted for. In FAIR we have Response and Productivity Loss. These losses are actual loss amounts accounted for in the analysis. When determining the pure premium in insurance, an actuary typically accounts for these types of expenses as factors applied to the losses. Some of these are known as Loss Adjustment Expenses (LAE) and account for such things as direct claims expenses (Allocated LAE) and more generalized claims expenses (Unallocated LAE).

When talking about losses, people are known to throw around terms such as "Once in Generation", "Hundred Year Storm" or similar phrases after a natural catastrophe. Usually, these are weather-related such as hurricanes or earthquakes. In the insurance realm, it is typically easy to know when they occur, since a government agency typically decares an event as a "Disaster". I was going to analyze some data to see how often a mega-breach (1 million records) occurs, thinking this would be akin to a natural catastrophe. However, while starting to look at the data, there are dissimilarities between the two realms that in my opinion require much more research.

- ✓ A hundred-year flood would affect many, whereas cyber breaches typically affect only one company. So, while we do have some predictive capabilities for hurricanes, I am unaware of any such tool for cyber attacks. A particular vulnerability might be known, however, whether it will be exploited, or how bad it will be is unknown.

- ✓ Zero-days, are the cyber version of an earthquake. I reason that as of now, we have no predictive capabilities for either.
- ✓ Insurance has very sophisticated models for estimating insured losses and has decades of loss experience. Cyber models are in their infancy and as I've previously stated, industrywide loss experience is non-existent.
- ✓ What might be catastrophic for one company can be very manageable for another. For example, a breach of 100 records for a small mom & pop shop might be catastrophic, but easily manageable for a large retailer with billions in sales. Some relationship between breach size and customers needs further analysis.
- ✓ What kind of loss is it? Most insurance property losses are commodity type losses; ie. a 1500 sqft house is worth $X, a 1986 car is worth $Y, etc. Cyber losses have more variables that can affect losses; i.e. if PII data is compromised, how detailed is it? Was the company following SOX , PCI-DSS or other required regulations?

The most important thing to remember is, *know your data*! Over the years, I've affectionately called the adjustments I've made to the data "Massaging". GIGO, Garbage In, Garbage Out is very true. In my experience massaging of the data should take a great deal of time, many times it is longer than the final analysis. This allows you to put full faith into your results and therefore the ability to stand behind any judgment used. Hopefully, your data sources are clean, but what do you do if you have two different sources that both code the same variable differently or have different values for the same data point. How will you map between one and the other? How will you handle outliers? Do all companies report data on the same basis, are they adjusted in some fashion?

These are some of the challenges I enjoy dealing with and make it very rewarding when all the puzzle pieces are ultimately aligned. Often this is an iterative process. Once you begin your analysis, questions about your data may arise causing you to go back and perform some more massaging. Ultimately your analysis flows smoothly due to the hard work endured at the beginning of the process.