10

Risks in Value-at-Risk

Standard Value-at-Risk (VaR) methodology pretends to superior identification of the risks of large market losses. Instead, it indulges the common fallacy of extrapolating from small losses. It needlessly sacrifices precision in big-picture forecasts and reads too much into minor details. We can do much better by tracking short-term volatility and allowing for uncertainty.

Most financial assets offer a much broader range of outcomes than pay-or-don’t-pay. Estimating all the probabilities directly is cumbersome and imprecise. Instead we formulate summary statistics like mean and standard deviation.

Risk analysts worry a lot about big losses, so they often supplement mean and standard deviation with measures of (lower) tail risk. That is where Value at Risk (VaR) comes in. The name appeals, as it suggests a definite cap on losses.

In fact, VaR is closer to a floor than a cap. There’s nothing definite about it. And if calculated in the standard way, it is grossly inferior to a host of other measures. Using standard VaR to tame financial risk is like using cigarette filters to tame cancer risk.

A Chance Encounter

’Twas a chance encounter that opened my eyes to VaR chance. I was walking down the street when a man called out of an alley, “Buddy, can you spare a million dollars?”
He was well dressed but had a dazed look in his eyes. “You must be a banker,” I said.

“Correction. I was a banker. What kind of banker asks for only a million dollars?”

“Point taken. So what are you now?”

“A financial risk analyst.”

“That’s still expensive. Look, if you’re selling magic beans, I’ll give you a cow.”

“Very funny. What I offer is far more valuable than giant bean stalks. I will tell your value at risk.”

“I already know. If I give you a million dollars, I risk it all.”

He shook his head. “No, you don’t understand. I invest your money in liquid markets. Every day the value wobbles. By studying that wobble I can divine the most you will lose most of the time. That’s the Value at Risk, or VaR for short.”

“I can divine the most I will lose at any time. It’s everything.”

“That’s the most most. VaR refers to the least most. It’s the maximum loss in all but q% of cases.”

“And the minimum loss in the rest?”

“Well, yes . . . . I rank observed losses for a year or two and find a threshold where approximately q% are worse.”

“So if I’m a soldier in a VaR-rated vest facing a hail of bullets, I expect only about q% to get through. How comforting.”

“Finance isn’t war. If clients can’t take a few hits they shouldn’t invest. Besides, they can set q wherever they want. That’s one of the beauties of VaR.”

“And if clients don’t know where to set it?”

“Then I set it at five. It is relatively more precise than most other VaR estimators.”

“What’s wrong with the rest?”

“Nearly all estimators are wrong. The observed averages will seldom match the expected values going forward. In finance, the average deviation is typically a significant fraction of what we’re trying to estimate. But it’s a lot less for 5% VaR than for, say, the 0.1% VaR.”

“Really?”

“Sure. To begin with, I need about 1,000 observations to even calculate an empirical 0.1-percentile. Really I want at least two or three observations below to keep some extreme outlier from polluting the results. That’s about eight to 12 years of daily data, and I don’t know how much of the
older stuff will stay relevant. But even if I had centuries of relevant data, the far tails are relatively harder to pin down than the near tails.”

“Why is that?”

“Far tails focus on a small subset of outcomes. Also, VaR is crude. In a way it likens everything to default risk, with VaR the cutoff between default and servicing. Since the vast majority of observations exceed the VaR, most valuable information comes from the few that don’t.”

“Hmmm, reminds me of the problems in estimating tiny default rates. But even the 5-percentile you prefer will discard a lot of information. How does it compare with the ordinary standard deviation estimator?”

“It’s usually two or three times as crude.”

“So why not estimate the standard deviation and infer the tails from it?”

“That’s a normal mindset. Markets don’t have to be normal. Usually they’re not.”

“We don’t have to assume normal tails. We can assume a fat-tailed distribution and infer the tails for it.”

“Aha. Assumptions again. Empirical VaR doesn’t make those assumptions. It just measures the tails and uses them as forecasts. If the 5-percentile doesn’t give enough buffer, double it. That’s what a lot of my best clients do.”

“What if they triple or quadruple the empirical standard deviation instead?”

“They’d likely get better accuracy with less data. But that wouldn’t be empirical VaR, would it?”

“No, I guess not. But given what you’ve just revealed, do you think I’m a likely buyer?”

“No, I guess not.”

Fat Tails

When risks are normal, the chances are 5% that losses exceed 1.6 standard deviations, 1% that losses exceed 2.3 standard deviations, and 0.1% that losses exceed 3.1 standard deviations. Risks that exceed these levels are called fat-tailed. Most portfolio risks are fat-tailed.

Kurtosis, a measure of the standardized fourth moment, is the best single measure of tail fatness. It’s zero for a normal distribution and typically positive for fat tails. Indeed, a positive kurtosis is usually taken as the rigorous definition of fat tails.
As we saw with exchangeable debt portfolios, kurtosis can’t fully describe tail risks. That’s the main reason we might want to estimate tail risks directly. However, a measure at one tail doesn’t confine the conditional risk further out. In principle, the risk distribution can take any non-negative shape.

For benchmark analysis, it’s helpful to posit a distributional form that inflates normal tails in a smooth way. The most favored form is “Student’s $t$,” also known as the $t$-distribution. In addition to mean and variance, it depends on a positive parameter known as the degree of freedom. However, to facilitate comparisons I will express everything in terms of kurtosis. This works as long as kurtosis is finite.

Student’s $t$ arises naturally in the estimation of confidence intervals. It amounts to a mixture of normal distributions of different variances, with the inverse variances following a gamma distribution. As is evident from Figure 10.1, Student’s $t$ densities decay log-quadratically near their peaks before slowing to log-linear or less. The only exception is the zero-kurtosis case, which corresponds to Gaussian normality and decays log-quadratically throughout.

To maintain the same standard deviation as a thin-tailed risk, a fat-tailed risk must have a taller head. That is, it must pile up more density near

![Figure 10.1](https://cupola.columbia.edu)

Log Densities of Student’s $t$-Distributions
the center. The cumulative distributions must match somewhere in the shoulders. For $t$-distributions with finite kurtosis, that somewhere is around 2 standard deviations, near the 2.5-percentile. (This helps explain the widespread association of 95% confidence with a $\pm 2$ standard deviation confidence interval.)

The 5-percentiles and 1-percentiles aren’t far apart, either. In standard deviation terms, the 5-percentile ranges from 1.5 to 1.6; the 1-percentile ranges from 2.3 to 2.6. The spans are around a tenth of the midpoint.

The differences widen farther out into the tails. The 0.1-percentile ranges from 3.2 to 5.1 standard deviations. The span is nearly half of the midpoint. If we know only the 5-percentile, converting it to a 0.1-percentile might require less than a doubling or more than a tripling.

Hence, even when tails scale smoothly, it’s dangerous to infer the outer tail risks from either the standard deviation or the inner tails. Empirical VaR seeks to estimate the tail risks directly. If we take 999 observations and rank them from lowest to highest, we might reasonably assign the lowest to the 0.1-percentile, the next lowest to the 0.2-percentile, and so on. If we have more observations and their assignments skip over the percentile we’re interested in, we interpolate.

Since empirical VaR makes no assumptions about distributional form, it might appear quite robust. It concentrates on what actually happened, without using theory as a filter. However, this is also its critical weakness. It reads too much into accident. As an estimator, it’s far less precise than its name suggests.

Imprecision

Let’s get more precise about imprecision. The simplest measure is the mean squared error. If on average the estimator gets the mean value right, the mean squared error will on average match the variance, making it the inverse of the Bayesian precision we met in Chapter 7. That’s appealing.

However, if we rank risk estimators by mean squared error, estimators of tiny risk are bound to look better than estimators of huge risk. Usually we care far more about the percentage error. The standard relative error (SRE) is the square root of the mean squared percentage error. The SRE is never less than the percentage bias (the relative distortion of the true mean by the estimator) or the percentage standard deviation, and never more than their sum.
As the standard deviation scales inversely to the square root of the number $T$ of independent observations, we’ll always prefer nearly unbiased estimators in sufficiently large samples. However, in analyzing daily portfolio returns, we rarely have more than a few hundred observations we’re confident are relevant and independent. This forces a trade-off between what we’d ideally like to estimate and what we can estimate best.

For example, suppose we estimate the standard deviation of a market portfolio as the square root of the mean squared daily return. This is a biased estimator since it ignores both the sample mean return and the distortion imposed by taking a square root. However, for most markets and at least 20 independent observations, the bias will be small. The imprecision works out to

$$SRE(\text{ordinary std dev}) \equiv \sqrt{\frac{2 + \text{kurtosis}}{4T}}.$$  \hspace{1cm} (10.1)

To achieve an SRE of 10%, we’ll need roughly 50 observations when kurtosis is 0 and 200 when kurtosis is 6. For fat tails and moderate numbers of observations, other estimators of standard deviation work better. The simplest multiplies the mean absolute return by 1.3. While often biased by a few percentage points, it dampens sensitivity to extreme returns. Variants on this are good foundations for more complex estimators.

Turning to percentiles, Kendall and Stuart (1972) provide a large-$T$ approximation that can be expressed as

$$SRE(q\%-ile) \equiv \frac{\sqrt{q\%(1-q\%) / T}}{\text{density}(\text{threshold}) \cdot \lvert \text{threshold} \rvert}.$$  \hspace{1cm} (10.2)

To achieve an SRE of 10% with a 5-percentile and $t$-distributed risk, we’ll need over 150 observations when kurtosis is 0 and nearly 300 observations when kurtosis is 6. That’s a lot less efficient than the ordinary standard deviation, although the relative gap narrows as tails fatten. While the 5-percentile can achieve a better SRE than equation (10.1) for a kurtosis exceeding 12, even then it will tend to be much less efficient than a trimmed standard deviation.

The SRE at other percentiles is worse, and soars for extreme tails. To achieve an SRE of 10% with a 0.1-percentile and $t$-distributed risk, we’ll need about 1,000 observations when kurtosis is 0 and 9,000 observations when kurtosis is 6. As a result, financial analysts rarely rely on direct esti-
mation of the 0.1-percentile. Instead they infer it from data closer to the center. This belies the oft-made claims that VaR measures extreme risks without assumptions on the distribution.

In general, when the far tails are easy to estimate directly, we don’t care that much about doing so. When we do care about them they’re hard to estimate. Standard VaR promises something it can’t deliver.

VaR’s poor precision reflects its focus on a sliver of information. The vast majority of observations simply confirm the noncriticality we expected. For most of the rest, the estimator cares only about which side of a threshold they fall on. The distance from the threshold matters only if it’s small enough for a change in threshold to flip sides. Contrast that with our standard deviation estimators, which incorporate every magnitude.

As Jorion (1997) notes, we’re better off inferring VaR from the standard deviation than using the empirical VaR. Granted, we may need additional information, like an appropriate kurtosis range. Zumbach (2007) presents empirical evidence for modeling risk with $t$-distributions having kurtosis between 1.5 and 12. RiskMetrics, the best-known commercial provider of risk analysis tools, revised its methodology to incorporate this insight. Unfortunately, there is scant regulatory pressure to enforce high standards of analysis.

Worse, too few risk managers appreciate the difference. For example, it is widely believed that, whatever its other shortcomings, standard VaR better captures extreme tails than standard deviation does. This is untrue. For $t$-distributed risk, the 5-percentile gets smaller in standard deviation terms as the extreme percentiles grow. This reflects the crisscross we noted earlier between densities having the same mean and standard deviation.

A Practical Illustration

In practice, neither volatility nor kurtosis stays constant, and a single $t$-distribution underestimates the rockiness of risk. The very coarseness of empirical VaR can help it stay robust. However, practice brings another problem to the fore. The standard rolling window of observation is typically too long to stay relevant.

Many risk managers think the opposite. They believe that averaging over multiple years is bound to improve precision. They forget that what matters most for forecasting is the identification of the current regime and its likelihood of switching, not some average over past regimes.
This is true whether or not one is investing for the long haul. Suppose we own an orchard facing severe frost in January. Would we justify a failure to turn on heaters by noting that it’s plenty hot in July and that we want the trees to last for years? Of course not. Then why use a thermometer measuring long-term averages, regardless of the evidence the weather has changed?

For a practical illustration, let’s see how various VaR and standard deviation metrics would have forecast the risks of the S&P 500 Index (SPX). For each day from January 1954 through August 2010, I compared the actual log return with the risk threshold estimated for it. The estimation used the previous $T$ daily returns as a rolling observation “window.”

The first metric was a 0.1-percentile estimated on a four-year window. The limit was breached 44 times, versus an expected value of 15 if the indicator were accurate. The chance that occurred strictly through bad luck is one in a billion.

To check for overcautious risk limits, I reviewed every quarter of every year for losses that exceeded half their threshold. If none occurred, I rated the quarter unusually calm. Whenever two consecutive quarters were unusually calm, I scored this as one “under limit” event.

Half of a 0.1% threshold is typically between a 5% threshold and a 2.5% threshold, implying a 4% to 20% chance of unusual calm. With 225 quarters examined, fewer than 10 under-limit instances are expected, unless tails are unusually fat or the estimated threshold was too high. The actual number was 38.

I then proceeded to convert other estimators into a 0.1-percentile proxy and test their forecasts similarly. I multiplied the 0.1% estimator by 1.5, the 5% estimator by 2.5, and the standard deviation estimator by -4. These multipliers represent the averages for $t$-distributions of kurtosis 1 to 3, rounded to the nearest 0.5. I refrained from finer calibration to reduce over-fitting.

The results broadly confirmed the superiority both of standard VaR projections over direct estimation of small percentiles and of standard deviation over standard VaR. See Table 10.1. However, the differences using like windows were less than simple theory suggests. What most stands out is the superiority of short windows. One-year windows fed better predictions than two-year windows. Six-month windows fed better predictors than one-year windows. The best predictor of the lot was a rolling three-month standard deviation.
If risks are stable, the ordering should be reversed. So risks must be fluctuating on time scales of a few months.

Figure 10.2 confirms that. The dotted line is a proxy for realized SPX volatility from January 2000 to August 2010. I constructed it using a trimmed standard deviation on a window stretching one month forward from the date in question.

With a window that compressed, the SRE is at least 15%. That accounts for a jiggle over 0.5 logarithmic units wide. However, the jiggle can’t explain the big secular changes. The total variation is 2.4 in log terms, or a factor of 12 in nominal terms. It took less than two years to go from trough in December 2006 to peak in October 2009, and most of that was reversed within a year.

<table>
<thead>
<tr>
<th>Type</th>
<th>Percentile</th>
<th>Multiplier</th>
<th>Window Length</th>
<th>No. Over Limit</th>
<th>No. Under Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>VaR</td>
<td>0.1</td>
<td>1</td>
<td>4 years</td>
<td>44</td>
<td>38</td>
</tr>
<tr>
<td>VaR</td>
<td>1</td>
<td>1.5</td>
<td>2 years</td>
<td>46</td>
<td>16</td>
</tr>
<tr>
<td>VaR</td>
<td>1</td>
<td>1.5</td>
<td>1 year</td>
<td>42</td>
<td>8</td>
</tr>
<tr>
<td>VaR</td>
<td>1</td>
<td>1.5</td>
<td>6 months</td>
<td>48</td>
<td>1</td>
</tr>
<tr>
<td>VaR</td>
<td>5</td>
<td>2.5</td>
<td>2 years</td>
<td>41</td>
<td>34</td>
</tr>
<tr>
<td>VaR</td>
<td>5</td>
<td>2.5</td>
<td>1 year</td>
<td>33</td>
<td>13</td>
</tr>
<tr>
<td>VaR</td>
<td>5</td>
<td>2.5</td>
<td>6 months</td>
<td>36</td>
<td>2</td>
</tr>
<tr>
<td>VaR</td>
<td>5</td>
<td>2.5</td>
<td>3 months</td>
<td>39</td>
<td>0</td>
</tr>
<tr>
<td>Std Dev</td>
<td>−4</td>
<td></td>
<td>2 years</td>
<td>39</td>
<td>34</td>
</tr>
<tr>
<td>Std Dev</td>
<td>−4</td>
<td></td>
<td>1 year</td>
<td>31</td>
<td>13</td>
</tr>
<tr>
<td>Std Dev</td>
<td>−4</td>
<td></td>
<td>6 months</td>
<td>28</td>
<td>4</td>
</tr>
<tr>
<td>Std Dev</td>
<td>−4</td>
<td></td>
<td>3 months</td>
<td>27</td>
<td>2</td>
</tr>
<tr>
<td>Std Dev</td>
<td>−4</td>
<td></td>
<td>1 month</td>
<td>30</td>
<td>1</td>
</tr>
</tbody>
</table>
No indicator that equally weights a year or more of data can keep up with that pace of change. It will lag six months. An indicator that equally weights three months of data will lag only six and a half weeks. The difference is quite noticeable in the chart, which overlays volatility predictors based on 1-year 5-percentile VaR (I divided by 1.6 before annualizing) and three-month standard deviation.

Sharp fluctuations in realized volatility explain why the three-month 5-percentile (which basically takes the third lowest observation) outperforms the two-year standard deviation. The latter has a colossal advantage in static efficiency. Yet that becomes irrelevant when the risk regime changes so much so quickly.

Figure 10.2 also reveals one of the main attractions of standard VaR: it’s relatively stable. Using a rolling window, over 90% of new evidence leaves the estimator unchanged. This naturally focuses VaR reports on the subset of changes. Other estimators hop around more; their reports require more work to filter out significant changes from noise.

The flip side is that VaR can get stuck in a rut. The chart displays an extreme example in 2009. For nine months the one-year 5-percentile held rock steady even though short-term volatility more than halved. During that period, standard VaR was useful only in predicting itself.
Scoring Rules

A short lag isn’t always better. When risk is relatively stable with only a mild trend, as occurred between 2004 and 2006, a longer-term measure will track it with less noise. When risk oscillates, as it did between 2000 and 2002, a short-term measure can get out of phase.

For simpler updating rules and more plausible motivation in terms of regime switching, we can substitute EMAs for simple moving averages. However, EMAs will still respond fast or slowly depending on their weighted-average lag, with the same basic trade-off between reliability and relevance.

Since the optimal trade-off can change without warning, we’ll usually get better results with a dynamic mixture of short-, medium-, and long-term measures. We can view the relative weights \( \{ p_i \} \) as probabilistic convictions that we update using Bayes’ Rule. A simple approximation that generally works well is

\[
\Delta p_i \equiv \frac{1}{3} p_i (\text{forecast}_i - \text{mean(forecast)}) - \frac{\text{var(news)}}{\text{var(news)}}.
\]  

(10.3)

Elsewhere I have called this approach SAMURAI (Osband 2002–2005). That’s short for Self-Adjusting Mixtures Using Recursive Artificial Intelligence. It slices through a lot of estimation knots.

Remarkably, SAMURAI can also be viewed as a market game, in which each forecaster has credibility capital \( w_i \) and bets near-optimally. “Near-optimal” means using the equivalent of a fractional Kelly criterion (Kelly 1956; Thorp 2000). Unlike much of finance, a fractional Kelly criterion is something both theorists and lauded practitioners tend to agree on. For an excellent popular account, see Poundstone (2005).

Other predictors draw on the prices of market options. They’re usually superior to crude historical statistics. Some people say that’s because markets look forward while statistics look backward. But that grossly overstates the difference. All predictions look backward into the future. The market’s advantages lie in mixing different predictors and infusing human intuition. From a machine perspective, human brains’ superb filters for relevance often outweigh their sloppy recall and crude calculation.

Whichever forecasting methods we choose, we should have ways of measuring their performance so that we learn what works best and don’t encourage intentional misreporting (Osband and Reichelstein 1985). This
is a classic problem in both weather forecasting and economic planning, and has stimulated much research on “proper” scoring rules. “Proper” means that the expected score is never higher for a dishonest forecast than for an honest one. See Gneiting (2010) or Gneiting and Raftery (2007) for recent surveys.

My heuristic comparison of SPX risk outliers doesn’t qualify because it doesn’t award a single aggregate score. An SRE doesn’t qualify because it doesn’t encourage unbiased estimation. However, proper scoring rules are easy to find for moments or percentiles, as Savage (1971) and Thomson (1979) showed. Here are two of the simplest:

1. If a forecaster reports a $q$-percentile of $y$, score the outcome $x$ as $yq\%$ less any loss $y - x$ beyond the threshold.
2. If the forecaster reports an $m$th moment of $y$, score the outcome as $y(2x^m - y)$.

To test VaR more systematically, I assembled ten years of daily data for hundreds of different equity indices, commodities, foreign exchange swap spreads, and individual U.S. and Japanese stocks. I scored forecasts every day for predictors of standard deviation, 5-percentile and 1-percentile. To reduce noise and prevent a few outliers from dominating results, I aggregated daily results into hundreds of thousands of multimonth games, and aggregated the games into a couple hundred tournaments.

Standard VaR did not win a single tournament. It didn’t win even on its 5-percentile home ground. Whatever we use to measure risk, let it not be standard VaR.

Even when VaR is right, it can be wrong. Imagine that VaR requires limiting the risk of $100$ million trading loss to 5%, when the real risk is 9%. Imagine a contract that is equally likely to halve the loss or double it (or triple or quadruple it). From a VaR perspective that makes the trading book safe again, even though it’s a bum deal for whoever underwrites it. While every risk measure can be gamed, VaR is particularly vulnerable because it cares only about the frequency of excess losses and not their intensity.

Regulators take note. If your charges claim to rely on standard VaR for risk control, they’re either fooling you or fooling themselves. But you don’t need to prescribe a particular non-VaR method. Just insist they keep tabs on their own methods and continually strive to improve.
“I feel sorry for the American cigarette companies,” said Prometheus. “They got in so much trouble over exaggerated claims for filters. They should have introduced a Cancer-at-Risk standard. It would explain the worst cancer the healthiest 95% of smokers would develop in two years.”

“That wouldn’t have saved them,” said Pandora. “The standards for peddling tobacco are much stricter than the standards for peddling financial carcinogens. In finance, let the sucker beware. Besides, you overlook the power of the name.”

“A rose by any other name would smell as sweet. . . .”

“That’s because a rose is directly material. Value is a perception.”

“Markets make value material, by trading money for real goods and services.”

“By that criterion VaR isn’t even value. It’s a threshold. Have financial institutions issue puts to cover the losses beyond their VaR. Now there would be value. But it could deviate a lot from the VaR.”

“If VaR is such rubbish, why not ban its use?” asked Prometheus. “Or make users post big health warnings, like ‘VaR is hazardous to your financial health.’”

Pandora smiled. “I like the health warning idea. But I fear it would get tucked into the long disclaimers that lawyers write to each other. Besides, every estimate of risk can be converted into a VaR measure given some auxiliary assumptions. There’s nothing wrong with that.”

“People think empirical VaR is better, because it doesn’t need conversion.”

“That’s another big misconception. The only thing empirical VaR is great at predicting is itself. Most days it doesn’t change at all.”

“I like Osband’s approach. Pick a measure of observed tail risk, set up a market contest to predict it, and use the market-clearing consensus as the predictor. But I’m not that impressed with the measures he uses. He would do much better looking at trading ranges.”