Why risk is so hard to measure*

Jon Danielsson
London School of Economics

Chen Zhou
Bank of the Netherlands and Erasmus University Rotterdam

February 2017

Abstract

This paper analyzes the reliability of standard approaches for financial risk analysis. We focus on the difference between value–at–risk and expected shortfall, their small sample properties, the scope for underreporting risk and how estimation can be improved. Overall, we find that risk forecasts are extremely uncertain at low sample sizes, with value–at–risk more accurate than expected shortfall. Value–at–risk is easily deliberately underreported without violating regulations and control mechanisms. Finally, we discuss the implications for academic research, practitioners and regulators, along with best practice suggestions.

Keywords: Reliability of risk measures, manipulation, value–at–risk, expected shortfall, finite sample properties, Basel III. JEL codes C10, C15, G18

*Corresponding author Jon Danielsson. We thank Dong Lou, Robert Macrae, Lerby Ergun and the Economic and Social Research Council (UK) [grant number: ES/K002309/1] for supporting the research. All errors are ours. Updated versions of this paper can be found on `www.RiskResearch.org` and the Webappendix for the paper is at `www.ModelsandRisk.org/VaR-and-ES`. Views expressed by Chen Zhou do not reflect the official positions of the Bank of The Netherlands.

1 Introduction

Much analysis, be it in the practitioner, academic or policy worlds, makes extensive use of statistical risk measures. Still, in spite of their prevalence, the performance of such risk measures in operational situations is surprisingly poorly understood. This is a concern since minor variations in model assumptions can lead to vastly different risk forecasts for the same portfolio, forecasts that are all equally plausible ex–ante. That is problematic for many applications, especially where the cost of type I and type II errors is not trivial.

Addressing this issue motivates our work here. We investigate the performance of the most commonly used measures of risk under typical usage scenarios, analyzing the most common practices in statistical market risk modeling, identify under what conditions they deliver reliable answers, when and how they fail to live up to expectations, how they can be manipulated and when they should not be used. Ultimately, we make recommendations to academics, practitioners and regulators on the use of statistical risk measures.

Statistical risk measures play a key role in financial decision–making, starting with Markowitz's (1952) mean variance–analysis. Volatility is the appropriate risk measure in the special case of normally distributed asset returns and for elliptically distributed returns in mean–variance analysis. However, we have known, at least since Fama (1963) and Mandelbrot (1963), that financial returns are fat tailed, implying that volatility is not the right way to measure risk except when warranted by specific applications.

In response, a number of risk measures that do no depend on an assumption of return normality have been proposed. The first of these, and the most prominent, is JP Morgan's 1993 value–at–risk, VaR. Since then a large number of alternatives have been suggested, however, the only one to get serious traction is expected shortfall, ES, (Artzner et al., 1999). VaR and ES have seen increasingly widespread use in decision–making, in no small part due to the Basel Committee choosing it as the measure of market risk in the Basel I accord in 1996. The Committee is now set to replace VaR by ES in Basel III; see the Basel Committee on Banking Supervision (2014).

The academic finance literature has made extensive use of statistical risk measures. Such work is usually based on the simplest risk measure of all, volatility. A number of papers relax the normality assumptions and explicitly use risk measures like VaR or ES. For example, Guptaa and Liang (2005) use VaR to calculate the capital adequacy for hedge funds and Patton (2009) uses

VaR as his definition of neutrality in his test of the neutrality of hedge funds. More broadly, Ibragimov et al. (2011) study diversification behavior under VaR constraints and Agarwal and Naik (2004) consider optimal portfolio construction under a mean–ES framework and Christoffersen et al. (2012) use ES for measuring dynamic diversification benefits in portfolios. ES has also been used as hedging criteria, e.g. by Cummins et al. (2004) who consider the catastrophic–loss index options to hedge hurricane losses. Furthermore, both VaR and ES have been used as the basic ingredient for constructing systemic risk measures as by Acharya et al. (2010), Brownlees and Engle (2015) and Adrian and Brunnermeier (2016).

The received wisdom maintains that VaR is inherently inferior to ES, a view supported by three convincing arguments. First, VaR is not a coherent measure unlike ES, as noted by Artzner et al. (1999). However, Daníelsson et al. (2012) show that VaR is also coherent provided the second moment of asset returns is defined. Second, as a quantile, VaR is unable to capture the risk in the tails beyond the specific probability, while ES accounts for all tail events. Finally, it is easier for financial institutions to manipulate VaR than ES. Here manipulation refers to cherrypicking certain risks that can lead to a situation in which either a given risk measure cannot reflect the chosen risk, or it is difficult to detect a misestimation, usually underestimation, of the given risk measure. Perhaps swayed by the theoretical advantages, ES appears increasingly preferred both by practitioners and regulators, most significantly as expressed by the Basel III Proposal. While the Proposal is light on motivation, the little that is stated only refers to theoretic tail advantages. The main theoretic disadvantage of ES over VaR is that while it is relatively straightforward to backtest VaR since we can compare predicted quantiles to actual observed returns, it is harder to backtest ES since one would need to compare a prediction of an expectation to some estimated expectation. The picture is not as clear cut once one looks at the practical implications. ES is estimated conditional on VaR and may therefore just increase the estimation error. Alternatively, because it smooths out the tails, it may be estimated more precisely.

Outside of academia, statistical risk measures are firmly embedded in financial regulations and are set to play an ever increasing role in how financial institutions operate. This applies now to the banking industry globally and most of the insurance industry. If the indications from the G20 and the Financial Stability Board are anything to go by, larger asset managers will be included soon. Therefore, the bulk of the financial sector either is required, or may be required to use statistical risk measures, typically VaR or ES, as a key ingredient in how they manage risk and determine capital. Once a rather specialist area, risk measures are of central importance to policymakers and industries alike and critical to decisions involving trillions of dollars worldwide, every year.

In spite of all of these applications, there are few studies of the practical properties of the risk measures. Two examples are Alexander and Sarabia (2012) who develop a framework for quantifying model risk, including the uncertainty in statistical risk measures, while Aussenegg and Miazhynskaia (2006) demonstrated via bootstrapping that the uncertainty in VaR estimation is considerable. Despite that, there is plenty of work on the theory, asymptotics and properties of particular estimators for risk measures, much less is known on what sort of estimation accuracy end users may expect. We surmise that an important reason relates to computational difficulties, we are estimating not only the risk measures but also the uncertainty of those estimates.

Our work is different from the many studies focussed on the performance of statistical methods. The performance of a given estimation method for risk measures is usually evaluated by checking whether the point estimate obtained from the method can predict large downside movement of the underlying time series in an out-of-sample exercise; see, e.g. Berkowitz and O'Brien (2002). Instead, we focus on the estimation uncertainty of a given method, i.e. not only the point estimate, but also the distribution of the potential estimated risk measure. In this regard, our study is not limited to a specific statistical method.

A large number of statistical methods for estimating VaR have been proposed, ranging from nonparametric historical simulation (HS) to highly sophisticated parametric methods. Amongst end users, there is a marked preference for more simple, and hence more easily implemented methods, and as a practical matter, only a handful have found significant traction, as discussed in Daníelsson et al. (2016). Of these, all but one depend on some parametric model, while one, HS, is model independent, which is what we opted for. It is a commonly used method, 60% of the US banks considered by O'Brien and Szerszen (2014) use HS. More fundamentally, the good performance of a specific parametric model is usually driven by the fact that the model is close to the data generating process (DGP) and it is not possible to find a parametric model that performs consistently well across all DGPs. Although HS is the simplest estimation method, it has the advantage of not being dependent on a particular parametric DGP, and any other method would be biased towards its particular DGP, creating an uneven playing field.

We address our questions of interest from both theoretic and empirical points of view. We start with investigating the uncertainty in the risk of stocks from the universe of stock returns in the Center for Research in Security Prices (CRSP) dataset. We follow that by theoretic and simulation analysis which allows us to study the properties of the risk measures when we know the DGP.

In our first contribution, we study whether the estimation of risk measures is robust when considering small — and typical in practical use — sample sizes. Although the asymptotic properties of risk measures can be established using statistical theories, known asymptotic properties of the risk forecast estimators might be very different in typical sample sizes such as the 250 daily observations used by the financial authorities in the Basel traffic light system for backtesting VaR 99%. We find that as the sample size gets smaller, the estimation uncertainty of both VaR and ES becomes extremely large. At the smallest samples, often the most commonly used in practice, the uncertainty is so large that the risk forecast can be as low as half the true value, or three times the true. In addition, the confidence interval around the risk forecasts is very far from being symmetric, the upper 99% confidence bound is a multiple of the forecast, which obviously cannot be the case for the lower confidence bound. This means that if one uses the standard error as a measure of uncertainty, it will be strongly biased downwards.

While many readers will not be all that surprised by the estimation uncertainty, we, at least, did not expect it to be as high as it turned out to be. Certainly, the implications do not seem to be well understood, even among policymakers and regulators with a specialist interest in this area. Other end–users, such as senior managers and regulators or non–expert users, often have a tendency to attribute a higher degree of faith in the risk measures than is warranted. Therefore, the fundamental and novel contribution of our work is the formal documentation of when risk measures work, how one should implement them, what one should be worry about, and what should be avoided.

We then investigate whether commonly used methods for evaluating the quality of risk estimates, both in–sample and out–of–sample, can be affected by the large estimation uncertainty. Focusing on the Basel traffic light approach and VaR, we find that for any entity subject to control by a risk measure, it is straightforward to underreport risk while remaining fully compliant with the letter of the control rule. Perversely, the entity is incentivized to pick the highest amount of tail risk because that allows it to maximize the amount of underreporting, and by violating the letter of the law, it might pick less tail risk.

In our third contribution, we investigate whether one can exploit the fact that non–extreme risk is estimated much more accurately than more extreme risk to create a scaling factor for risk across probabilities — probability shifting. That is indeed possible, at the expense of some uncertainty, giving the end user the opportunity to choose whether the estimation uncertainty is outweighed by the scaling inaccuracy.

We finally find that ES is estimated with more uncertainty than VaR. Since VaR and ES are in most cases related by a small constant, they are conceptually equally informative, and we find that in the special case of Basel III, the 97.5% ES is essentially the same as the 99% VaR. Even if ES is theoretically better at capturing the tails, in practice one might just multiply VaR by a small constant to get ES. There is however one reason to prefer ES, it is harder to manipulate.

The structure of the paper is as follows. We discuss the various properties of the risk measures in Section 2, the scope for underreporting risk in Section 3 and probability shifting in Section 4. We discuss the specific implications for financial regulations in the concluding Section 5. Mathematical proofs are relegated to the Appendix.

2 How accurate are risk measurements?

The statistical measurement of risk is firmly embedded in the academic finance literature, the operations of financial institutions and financial regulations. In spite of their prevalence, surprisingly little is known about the performance of risk measures, especially in common usage scenarios. It is that deficiency that motivates our work in this section.

To start with, consider the accuracy of estimating VaR and ES at two probability levels, 90% and 99% for the CRSP universe of stocks across three sample sizes, 300 days, 1,000 days and 5,000 days or 20 years. While we do not know the true value of the risk measures, and so cannot directly validate the estimation uncertainty, it is straightforward to approximate it by a block bootstrapping procedure. We used the universe of CRSP stocks from 1926 to 2014 and report the results in Table 1, reporting uncertainty, both standard error and 99% confidence bonds around a risk measure normalized to one, where we use the HS estimation procedure. The details of the sample selection procedure and the bootstrap is given in the footnote to the Table.

These results came as a surprise to us. At the lowest sample sizes, 300 days, the estimation uncertainty is considerable. For example, we cannot

Table 1: The accuracy of risk estimation in CRSP stocks.

Note: The table shows the accuracy of estimating VaR and ES at two probability levels, 90% and 99% using daily returns on all liquid traded stocks on NASDAQ, NYSE or AMSE from 1926 to 2014. A stock is included if on the first day of the sample the stock has a share price above 5$, its market capitalization exceeds the 10% quantile of the market capitalization of all stocks traded on NYSE on that day, and has 600 or more observations. Each stock return series is split into non–overlapping samples with sample sizes $N = 300, 1,000, 5,000$. The samples from different stocks are pooled together. The risk measure is estimated for each sample. In addition a bootstrapped sample is constructed by bootstrapping N/B blocks with a block size B. The standard error of the risk estimate for a given sample is calculated by the standard deviation of K bootstrapped estimates, while the 99% confidence interval is given by the 0.5% and 99.5% quantiles of the K bootstrapped estimates. The reported standard errors and the upper and lower bound of the 99% confidence intervals are the average across all samples.

			VaR		ES	
N	p	Samples	standard error	99% conf. bound	standard error	99% conf. bound
300	1%	77,669	(0.21)	[0.65,1.49]	(0.16)	[0.63,1.28]
1,000	1%	21,188	(0.13)	[0.74,1.35]	(0.14)	[0.69,1.35]
5,000	1%	2,538	(0.07)	[0.84,1.20]	(0.09)	[0.80,1.24]
300	10%	77,669	(0.12)	[0.75,1.30]	(0.11)	[0.74,1.29]
1,000	10%	21,188	(0.08)	[0.81,1.23]	(0.09)	[0.80,1.24]
5,000	10%	2,538	(0.05)	[0.89,1.13]	(0.05)	[0.87,1.15]

discriminate between for VaR 99% in the range of 0.65 to 1.49. When the sample size hits 20 years, or 5000 observations, the VaR 99% confidence bound is still between 0.84 and 1.2. Since as a practical matter, we are more likely to be closer to 300 days than 5,000 days in most cases, the estimation uncertainty is considerable.

Furthermore, ES is estimated with more uncertainty than VaR for the two longer sample sizes. This results contradict the prevailing wisdom that ES is superior to VaR. However, such arguments are based on the theoretical properties of the two risk measures, while our results come from finite sample empirical results.

Finally, the estimation uncertainty problem is mitigated when considering lower level probability, such as the 90% shown in the Table, where the reduction in estimation uncertainty is most pronounced for the lowest sample size, 300. This will be important to our proposal in Section 4.

This leaves the question of what is behind these results. Do they arise because of the particular stocks we chose and the way we did the bootstrap procedure, or is this something more general at work here? To answer the question, we employed both theoretical analysis of the statistical properties of risk measures as well as a Monte Carlo procedure.

2.1 Theoretic analysis

The empirical CRSP results above are confirmed by theoretic analysis. To start with, denote the profit and loss of a trading portfolio as PL and let $X \equiv -PL$, so we can indicate a loss by a positive number. Suppose X follows a distribution function F and obtain a sample of size N from that distribution. Indicating a particular probability level by p, denote VaR by $q_F := q_F(p)$.

Consider the best case scenario where the data is i.i.d. and we know it is i.i.d. If we also had to estimate the dynamic structure, the estimation uncertainty would be further increased. In Extreme Value Theory (EVT), heavy–tailedness is defined by regular variation in the tail of F:

$$\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = x^{-\alpha},$$

where $\alpha > 0$ is known as the tail index. For the student–t distribution, the tail index equals the degrees of freedom, with the Gaussian as the special case where $\alpha = +\infty$. Note that the assumption of regular variation only applies to the right tail of F and thus does not impose any restriction on the rest of the distribution, allowing this approach to capture a large range of models. An assumption of regular variation in the right tail is sufficient for inference on the distribution of tail risk measures.

Since we are dealing with tail quantiles with p close to 1, we employ the asymptotic theory for tail quantiles as in Einmahl (1992).[1] Theoretically, we consider an intermediate sequence $k_q := k_q(N)$ such that $k_q \to \infty$, $k_q/N \to 0$ as $N \to \infty$, and then investigate $\hat{q}_F$ with a probability level $1 - k_q/N$. By ranking the N observations $X_1, \cdots, X_N$ as $X_{N,1} \leq X_{N,2} \leq \cdots \leq X_{N,N}$, the HS estimate for the VaR is

$$\hat{q}_F(1 - k_q/N) = X_{N,N-k_q}.$$

[1] The general asymptotic theories of empirical quantile estimators are known, see, e.g. Theorem 6.2.1 in Csörgő and Horváth (1993).

The following proposition gives the asymptotic properties of the estimator $\hat{q}_F(1-k/N)$ for a general k sequence under heavy–tails. The proof is postponed to the Appendix.

Proposition 1 *Suppose $X_1, \cdots, X_N$ are i.i.d. and drawn from a heavy tailed distribution function F with $\alpha > 2$. Denote $U = \left({}^{1}\!/_{(1-F)}\right)^{\leftarrow}$ as the quantile function. Then ${}^{U(tx)}\!/_{U(t)} \to x^{-1/\alpha}$ as $t \to \infty$. Assume the usual second order condition that quantifies the speed of convergence in this limit relation (see, e.g. Condition (2.3.22) in de Haan and Ferreira (2006)):*

$$\lim_{t\to\infty} \frac{\frac{U(tx)}{U(t)} - x^{1/\alpha}}{A(t)} = x^{1/\alpha}\frac{x^\rho - 1}{\rho},$$

for a constant $\rho \leq 0$ and a function $A(t)$ such that $\lim_{t\to\infty} A(t) = 0$. Suppose $k := k(N)$ is an intermediate sequence such that as $N \to \infty$, $k \to \infty$, $k/N \to 0$ and $\sqrt{k}A(N/k) \to \lambda$ with a constant λ. Then, we have that as $N \to \infty$,

$$\sqrt{k}\left(\frac{\hat{q}_F(1-k/N)}{q_F(1-k/N)} - 1\right) \xrightarrow{d} N\left(0, \frac{1}{\alpha^2}\right).$$

Proposition 1 shows that the HS VaR estimator is asymptotically unbiased and provides a way to approximate its asymptotic variance as ${}^{1}\!/_{k\alpha^2}$. The thinner the tail (i.e. the higher α), the more accurate the VaR estimate.

The estimation uncertainty has a quantifiable economic impact. Given the true value of the quantile $q_F(1-k/N)$, Proposition 1 implies that the lowest possible estimate $\hat{q}_F(1-k/N)$ can be as low as

$$q_F(1-k/N)(1 - \frac{1}{\sqrt{k}}\frac{1}{\alpha}\Phi^{-1}(1-(1-\tau)/2)),$$

without being rejected under the confidence level $\tau < 1$.

The true value of the quantile $q_F(1-k/N)$ ensures, by definition, that $\Pr(X > q_F(1-k/N)) = k/N$ We now evaluate the probability that X is above the lowest possible estimate, and compare that to the intended probability k/N. From the heavy–tailed property, we get that

$$\begin{aligned}
&\frac{\Pr(X > q_F(1-k/N)(1 - \frac{1}{\sqrt{k}}\frac{1}{\alpha}\Phi^{-1}(1-(1-\tau)/2)))}{k/N} \\
\approx&(1 - \frac{1}{\sqrt{k}}\frac{1}{\alpha}\Phi^{-1}(1-(1-\tau)/2))^{-\alpha} \\
\approx&1 + \frac{1}{\sqrt{k}}\Phi^{-1}(1-(1-\tau)/2)
\end{aligned}$$

Putting this calculation in real perspective, consider the case for estimating VaR(99%) with a sample size $N = 1000$. k is then 10. Suppose we consider a typical confidence level such as $\tau = 95\%$. Then the ratio calculated is at $1 + {}^{1}/\sqrt{10} \times 1.96 = 1.62$. In other words, while we aim to get a VaR such that the probability of exceeding the VaR is 1%, with a reasonable estimate (not significantly different from the true value under 95% level of confidence), the probability of exceeding the estimated VaR can be as high as 1.62%. In a relative comparison, this is more than 60% higher.

Using ES does not help in reducing the estimation uncertainty. Denote ES at probability p by $e_F := e_F(p)$. Like we did for VaR, since we are dealing with p close to one, we can use an intermediate sequence k_e such that $k_e \to \infty$ and $k_e/N \to 0$ as $N \to \infty$. We then estimate ES at the level $1 - k_e/N$ as:

$$\hat{e}_F(1 - k_e/N) = \frac{1}{k_e} \sum_{j=1}^{k_e} X_{N,N-j+1}.$$

We have similar theoretic results for ES as we did for VaR.

Proposition 2 *Suppose $\alpha > 2$. Under the same conditions as in Proposition 1, we get that as $N \to \infty$,*

$$\sqrt{k}\left(\frac{\hat{e}_F(1-k/N)}{e_F(1-k/N)} - 1\right) \overset{d}{\to} N\left(0, \frac{2(\alpha-1)}{\alpha^2(\alpha-2)}\right).$$

By applying Propositions 1 and 2, we can compare the accuracy of VaR and ES as in the following corollary.

Corollary 3 *Assume the same conditions as in Proposition 2. Consider* VaR *and* ES *at the same probability level p. Then, as $N \to \infty$,*

$$\frac{\operatorname{Var}\left(\frac{\hat{e}_F(p)}{e_F(p)}\right)}{\operatorname{Var}\left(\frac{\hat{q}_F(p)}{q_F(p)}\right)} \to \frac{2(\alpha-1)}{\alpha-2}.$$

In the specific case of the Basel II and III risk measures, with VaR *at probability level p_1 and* ES *as probability level p_2 such that $(1-p_2)/(1-p_1) = 2.5$. Then, as $N \to \infty$,*

$$\frac{\operatorname{Var}\left(\frac{\hat{e}_F(p_2)}{e_F(p_2)}\right)}{\operatorname{Var}\left(\frac{\hat{q}_F(p_1)}{q_F(p_1)}\right)} \to \frac{4(\alpha-1)}{5(\alpha-2)} =: g(\alpha).$$

The corollary shows that the estimation inaccuracy of ES is higher than that of VaR. When considering the same probability level, since ${}^{2(\alpha-1)}/_{(\alpha-2)} > 2$, the relative estimation variance of ES is at least twice larger than that of VaR. When considering different probability levels, like VaR(99%) and ES(97.5%), and noting that the function $g(\alpha)$ is decreasing in α and the break–even point for $g(\alpha) = 1$ is reached at $\alpha^{\mathrm{be}} = 6$. This means that for $\alpha < 6$, the estimation uncertainty in ES(97.5%) is higher than that of VaR(99%).

We estimated α for the stocks in our CRSP and find that the mean $\alpha = 3.31$, and for all stocks, $1.71 < \alpha < 5.78$, and therefore the empirical result in Table 1 that VaR is better estimated than ES s is not all that surprising. It is what theory predicts.

2.2 Monte Carlo analysis

While the theoretic results above provide guidance as to the asymptotic performance of our risk measures, they are only asymptotic and do not tell us whether the asymptotic theory holds in typical sample sizes, which may range from a couple of hundred to a few thousand. For that reason it is of interest to investigate the properties of the risk estimates for a range of sample sizes that might be encountered in practical applications, such as those in the CRSP analysis above. We do that by means of an extensive simulation study.

We focus on i.i.d. observations drawn from the Student–t distribution, the i.i.d. Pareto distribution and a GARCH model with Student–t distributed innovations. The qualitative results from all three are similar, so we only present the Student–t results here, leaving the others to the web appendix. The sample size, N, ranges from 200 to 100,000, where for presentation purposes, we denote sample sizes above 300 as years with a year consisting of 250 observations. The results are generated for a range of probabilities. Such a procedure is repeated $S = 2 \times 10^7$ times, for a discussion on why this simulation size is needed see Appendix A.1. We then report the standard error of the VaR estimates relative to the true VaR, as well as the 99% empirical confidence interval. A representative subset of the results is shown in Table 2, with the full results in the web Appendix.

The results from Table 2 are in line with both the empirical CRSP and theoretic results above. The quality of the estimation increases as the sample size gets bigger, the tail thins and the probabilities fall. Furthermore, ES it is always estimated more inaccurately than VaR.

Table 2: VaR and ES 99% finite sample performance

Note: For each given α, N observations from a standard Student–t distribution with degree of freedom $\nu = \alpha$ are simulated. For each simulated sample, risk is estimated and then divided by its true value. The resulting ratio is regarded as the relative estimation error. The table reports the standard error and 0.5% and 99.5% quantiles of these ratios across the $S = 2 \times 10^7$ simulated samples.

			VaR		ES	
sample size	α	p	se	99% conf. interval	se	99% conf. interval
300 days	2.5	99%	0.33	[0.61,2.46]	0.56	[0.42,3.42]
300 days	2.5	90%	0.11	[0.76,1.34]	0.20	[0.68,1.76]
300 days	5	99%	0.18	[0.72,1.70]	0.22	[0.61,1.82]
300 days	5	90%	0.09	[0.79,1.27]	0.10	[0.78,1.29]
4 years	2.5	99%	0.15	[0.74,1.51]	0.31	[0.59,2.27]
4 years	2.5	90%	0.06	[0.86,1.17]	0.11	[0.80,1.39]
4 years	5	99%	0.09	[0.82,1.29]	0.12	[0.75,1.40]
4 years	5	90%	0.05	[0.88,1.14]	0.05	[0.87,1.15]
50 years	2.5	99%	0.04	[0.91,1.11]	0.09	[0.84,1.31]
50 years	2.5	90%	0.02	[0.96,1.04]	0.03	[0.93,1.10]
50 years	5	99%	0.02	[0.94,1.07]	0.03	[0.92,1.10]
50 years	5	90%	0.01	[0.96,1.04]	0.02	[0.96,1.04]

What is surprising, at least to us, is how large these changes are. With a tail index 2.5, probability 99% and sample size 300 days, the 99% confidence bound is [0.61,2.46] where the true value is one. It is highly asymmetric implying that if one would get misleading answers by using the standard error (0.33). If one goes to the other extreme with 50 years of data, tail index 5 and 90% probability, the VaR estimates are quite reasonable.

2.3 Intermediate conclusion

We believe these are the first results to comprehensively demonstrate the accuracy of statistical risk measures in typical applications. While a number of studies focus on comparing empirical methodologies of estimation procedures or theoretic properties of risk measures, our focus is on how risk estimation performs in typical usage scenarios.

Two main results emerge. First, the measurement of risk with a measure like VaR an ES is highly inaccurate, especially in the smaller sample sizes common in many applications, hundreds or perhaps a few thousand days. At 300 days, the 99% confidence bound for VaR is 0.65 to 1.49, in our sample of the universe of CRSP stocks, narrowing down to 0.84 to 1.20 when the sample size increases to 5000 days. This result extends across the empirical, theoretical and Monte Carlo analyses.

In our second result, we find that ES is generally estimated more inaccurately than VaR. This is at odds with the received wisdom in the regulatory, practitioner and academic domains where ES it is usually preferred. We surmise that this is because from a theoretical point of view, ES is superior to VaR. However, we demonstrate that this does not extend to practical applications. As a practical matter, when returns are fat tailed, ES is just a multiple of VaR, where the multiplication factor depends on the tail index. If ES is needed, the end–user might just as well estimate VaR and multiply it by a constant. Considering that it is much harder to backtest ES than VaR, there seems to be little reason to pick ES for most practical applications.

3 On the scope for underreporting risk

The measurement of risk, perhaps by VaR or ES as in the last Section, is not all that useful unless the end–user has some way of validating the quality of the risk measure. The problem facing those aiming to verify and control risk is that any mechanism used as the control device can skew incentives for taking risk, even if the intention is to fully comply with both the letter and the spirit of the control. For example, as we show below, in the case of the current market risk regulations, Basel II, the specific control mechanism perversely incentivizes banks to load up on tail risk. Even if the senior management of a bank does not intend to act in this way it is entirely possible that internal incentives (such as bonuses) may incentivize traders to do so. This has the effect of undermining the effectiveness of regulations, because of the impact noted by Perotti et al. (2011) where with sufficient tail risk, almost no level of capital is sufficient to protect an institution in times of failure.

This is complementary to the theoretical work of Cuoco and Liu (2006) who consider the optimal VaR reporting for banks. Our approach focuses on the room for potential underreporting. If a bank is able to measure the probability of detection, the bank can ex ante load up tail risk to achieve their desired underreporting room. Our results also echo Marshall and Prescott (2006) who find when the authorities control the variance of a portfolio, same as VaR control under normality, the banks try to increase the mean return by taking additional risk.

When it comes to validating risk forecast models, there are two main approaches one can take, observe model forecast performance in–sample or out–of–sample. Confusingly, the term backtesting can imply either approach, depending on who uses the term. Regulators, in particular, often use backtesting for out–of–sample analysis.

The banking regulation authorities, in Basel I, opted for an out–of–sample approach with the so–called "traffic light" system, based on how often actual portfolio losses exceed their VaR (violations) over a testing window. For example, to backtest VaR(99%), the traffic light system is based on the number of violations (l) in 250 days: if l is less than five the bank is in the green zone, if l is between five and nine the bank is in the yellow zone and the red zone for more violations. By the Bernoulli properties of violations, it follows that the probability (τ) of being in each zone, if the model is correctly specified, is 89.22%, 10.76% and 0.025%, respectively. In what follows, we consider a more general setup of backtesting VaR(p) with l violations in a testing window with size W.

Our results below apply only to the specific case of VaR not ES, as the former is based on a quantile, and hence straightforward to backtest, unlike ES where additional assumptions have to be made, since it is not elicitable (see e.g. Gneiting, 2011).

3.1 Risk underreporting room

From a bank's perspective, the specific backtesting mechanism may gives the bank the room to underreport the VaR while remaining compliant with the letter of the regulation. When backtesting a bank's reported VaR, the traffic light system allows for l violations in W days. Instead of viewing this as checking the number of violations, an alternative view is that, the regulator is estimating a VaR at a probability level $1 - l/W$ using the W observations in the testing window, and then compare the estimate with the number reported by the bank. Since the regulator would also suffer from estimation uncertainty, the bank can explore that by reporting a lower value while still retain the probability of a green zone. Mathematically, we calculate the ratio between the lowest possible value that can be reported by a bank and the true VaR, denoted as the *underreporting room* (UR). The UR is related to the estimation uncertainty issue raised in the last Section, because a bank can effectively pick a value that lies within the confidence band of the VaR estimate as the reported VaR.

Consider the Basel traffic light approach. The regulator chooses p, W and l. Suppose a bank focuses on the green zone, where the bank would have its own target of achieving the green light with probability τ and can choose the tail thickness of its portfolio α. Then the bank's UR depends on the five variables and is denoted by $\mathrm{UR}(\tau, \alpha; l, W, p)$, with the first two chosen by the bank and the last three set by the regulator.

Suppose a bank knows the values of the P/L in the testing window, X_t, for $t = 1, 2, \cdots, W$. Then the lowest possible value a bank may report while staying in the green zone is the $(l+1)^{\text{th}}$ highest order statistic, $X_{W-l,W}$ plus a small increment, set without loss of generality to zero. This gives exactly l violations. Ex ante, the order statistic is unknown, rather, only the distribution of X is known. Therefore, the bank has to study the property of the distribution of $X_{W-l,W}$ denoted as G.

Following Proposition 1, since l/W is a small positive number, the asymptotic distribution of $X_{W-l,W}$ is given as

$$\frac{X_{W-l,W}}{q_F(1-(l+1)/W)} - 1 \sim N\left(0, \frac{1}{(l+1)\alpha^2}\right).$$

Hence, we get that approximately, if the bank intends to get a green light with probability τ, the bank should report the τ-quantile of the distribution G as

$$q_F(1-(l+1)/W) \cdot \left(\Phi^{-1}(\tau)\frac{1}{\sqrt{l+1}\alpha} + 1\right),$$

where Φ is the standard normal distribution function. We compare this value to the true $\mathrm{VaR}(p)$. Note that

$$\frac{q_F(p)}{q_F(1-(l+1)/W)} \approx \left(\frac{(l+1)/W}{1-p}\right)^{1/\alpha} = \left(\frac{l+1}{W(1-p)}\right)^{1/\alpha},$$

which follows from the assumption of heavy tailed returns. This implies that the UR is

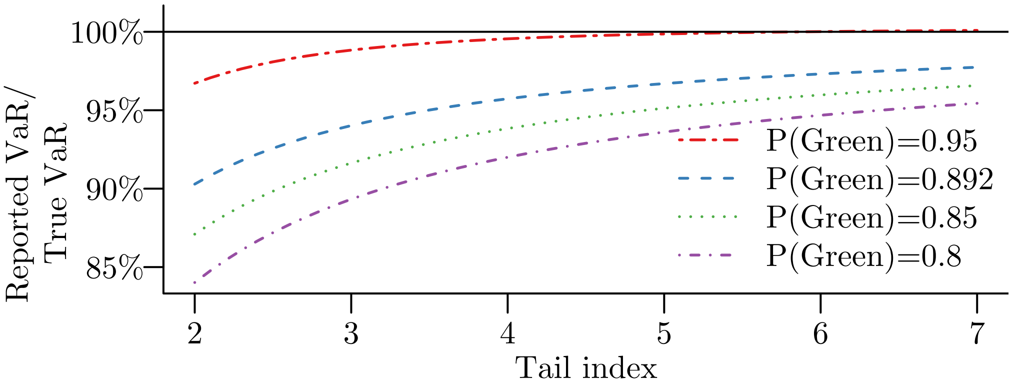
$$\begin{aligned}\mathrm{UR}(\tau, \alpha; l, W, p) :=& \frac{q_F(1-(l+1)/W) \cdot \left(\Phi^{-1}(\tau)\frac{1}{\sqrt{l+1}\alpha} + 1\right)}{q_F(p)} \\ \approx& \left(\frac{W(1-p)}{l+1}\right)^{1/\alpha} \left(\Phi^{-1}(\tau)\frac{1}{\sqrt{l+1}\alpha} + 1\right). \qquad (1)\end{aligned}$$

We show a typical UR ratio in Figure 1 for the green Basel zone. Then, $p = 0.99$, $W = 250$, $l = 4$, where the regulator targets $\tau = 0.8922$. The

Figure shows four cases, the first where the bank complies with the letter, but not necessarily the spirit, of the regulations ($\tau = 89.22\%$), two others where it deliberately does not ($\tau = 85\%$ and 80%), and finally one where it voluntarily takes a more conservative position ($\tau = 95\%$.) We capture this by calculating the $\mathrm{UR}(\tau, \alpha; 4, 250, 0.99)$ against α for four different levels of τ, 95%, 89.22%, 85% and 80%.

Figure 1: Underreporting room versus tail index

Note: The figure shows the UR ratio against the tail index (α) for four targeted levels of probability falling in the green zone (τ). The UR ratio is the ratio between the lowest possible value that can be reported by a bank and the true VaR defined in (1). Here the parameters are $W = 250$, $l = 4$ and $p = 0.99$. The dash (blue), dotted (green), dash-dotted (purple) and long dash-dotted (red) lines correspond to τ levels at 89.22%, 85%, 80% and 95% respectively.



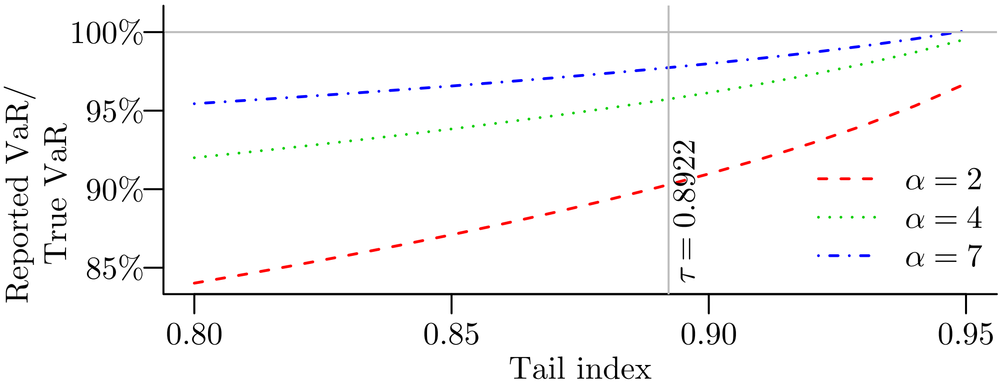
Two key results emerge from the figure. Even if the bank intends to comply with regulations, it has some room for underreporting risk, in the case of the typical $\alpha = 4$, by at least 5%. Also, the more heavy tailed its portfolio returns are, the more room the bank has to underreport risk. Only when a bank is seriously concerned about not obtaining a green light ($\tau = 0.95$), it may overreport the risk provided that its portfolio is not very heavy tailed.

In addition, we plot the UR ratio against various levels of τ ranging from 80% to 95% for three given levels of α, $\alpha = 2, 4$ and 7, in Figure 2. We observe that the underreporting room shrinks as the targeted probability of being in the green zone τ increases. Across different levels of asset tail risks, the underreporting room is more pronounced for heavier tails (lower α).

These results are directly affected by the size of the testing window. We further investigate this by increasing W. With a properly set l, we calculate how the UR is affected by sample size, reported in Figure 3. The UR is shrinking with increasing sample sizes, 250, 500 and 1,000, at $\tau = 90\%$. For $\alpha = 3$, the UR is 6% when the sample sizes one year, falling to 2% once the

Figure 2: Underreporting room versus targeted probability

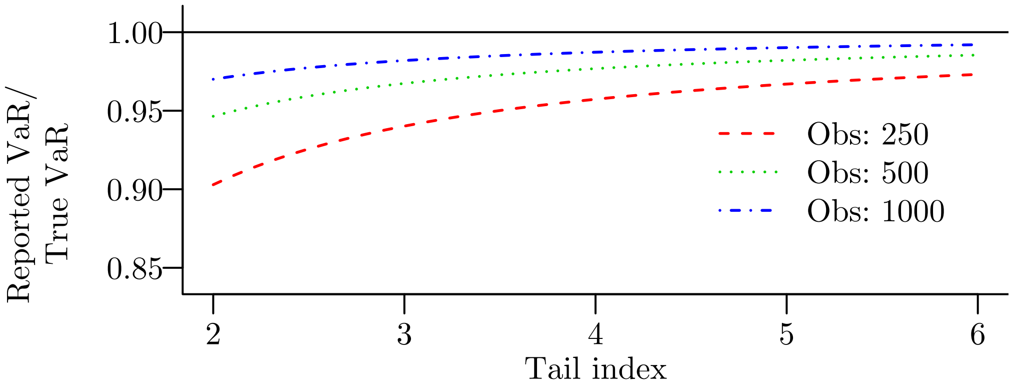
Note: The figure shows the UR ratio against the targeted levels of probability falling in the green zone (τ) for three levels of tail index (α). The UR ratio is the ratio between the lowest possible value that can be reported by a bank and the true VaR defined in (1). Here the parameters are $W = 250$, $l = 4$ and $p = 0.99$. The dash (red), dotted (green) and dash-dotted (blue) lines correspond to α levels at 2, 4 and 7 respectively.



testing window is four years.

Figure 3: Underreporting room with different sample size

The figure shows the UR ratio against the tail index (α) for different sizes of the testing window (W). Here the parameters are $p = 0.99$ and $\tau = 90\%$. The dash (red), dotted (green) and dash-dotted (blue) lines correspond to W levels at 250, 500 and 1000 respectively. The parameter l is accordingly adjusted to 90% quantile of a Bin(W,$1-p$) distribution.



Taken together, these results are especially problematic for out–of–sample model evaluation methods, like the Basel traffic light approach, suggesting that 250 days are not sufficient. Unfortunately, increasing that to a more accurate thousand days would leave the regulators to wait four years before they can pass judgment on the trading book risk models.

3.2 Risk taking under estimation uncertainty

What is more of a concern to us is the observation that the fatter the tails of the returns on the underlying trading portfolio, the more room a financial institution has to underreport risk. Consider a financial institution that is seen by the authorities as having too little trading book capital, or conversely too many risky assets. This institution may of course opt to raise capital or sell assets, but under current rules it has a third option because of the UR. In that case, suppose the bank underreports a fraction f of its risk. In what follows, we focus on the probability of not being in the red zone, but the same qualitative result obtains for the green zone.

Suppose the bank intends to achieve a probability of not being in the red zone light at level τ. Then there is an upper bound for the tail index the bank can accept, $\bar{\alpha} = \bar{\alpha}(f, \tau)$. In other words, if the bank has a portfolio with a tail index $\alpha > \bar{\alpha}$ and intends to achieve a probability of having a red light at level p, the corresponding UR is limited such that $\mathrm{UR}(\tau, \alpha; l, W, p) > f$. Therefore, we can use the inverse function of (1) to infer the maximum level $\bar{\alpha}$ that is acceptable. This can be achieved by solving $\mathrm{UR} = f$ with a given τ, l, W and p to obtain $\bar{\alpha}(f, \tau; l, W, p)$.

Figure 4 shows the value of $\bar{\alpha}(f, p; l, W)$ for $l = 9$, $W = 250$. We plot $\bar{\alpha}$ against against different level of τ for three level of f, $90\%, 85\%$ and 80%. The vertical line indicates the designed probability of having a red light in the traffic light system. In other words, if the bank fully intends to comply with the regulations, it sticks to that target probability.

The striking conclusion is that as the bank's actual level of compliance increases, the more tail risk it needs to introduce into its portfolio. This arises because if the bank has a fixed underreporting target, while not wanting to be detected by the test for having too much risk, the only remaining instrument at its disposal is to exploit the estimation uncertainty and increase tail risk.

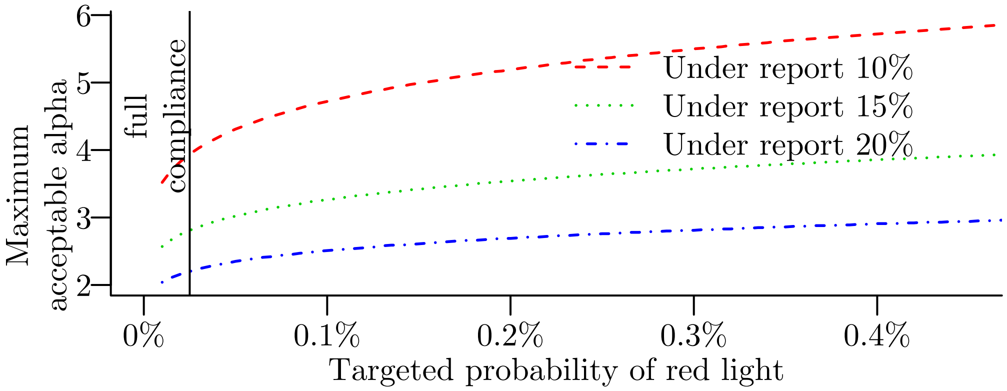
The perverse outcome is that the lower level of actual compliance is, the safer its portfolio becomes from the point of view of tail risk.

3.3 Implications

The main purpose of statistical risk measures in financial regulations and practical applications, is to control risk–taking. Financial regulators use such methods to set capital requirements for banks proprietary trading and

Figure 4: Maximum acceptable tail index for a given underreporting room

The figure shows the maximum acceptable tail index ($\bar{\alpha}$) against the targeting probability of not being in the red zone (τ), given a bank's intended UR. The dash (red), dotted (green) and dash-dotted (blue) lines correspond to three given UR levels (f) at 90%, 85% and 80% respectively. For each given f and τ, the $\bar{\alpha}$ is solved by inverting $\mathrm{UR}(\tau, \alpha; l, W, p) = f$, where the function UR is defined in (1). Here the parameters are $p = 0.99W = 250$ and $l = 9$.



internally, financial institutions use them for multiple purposes, including controlling traders, allocating trading capital between asset classes and compensating risk–takers. From the point of view of anyone intending to control risk–taking, our results are problematic.

In all of these cases, the incentive to take excessive risks ex ante is real as any entity subject to control by risk measures can either manipulate risk without changing the risk measure or taking specific tail risks to exploit the weakness in the control process. Even if senior management fully intends to obey the spirit as well as the letter of the rules it is entirely possible that internal incentives such as bonuses or promotions will lead to traders acting in this manner because maximizing risk is typically associated with maximizing profits.

There are several ways one can exploit the weakness in the control process. Perhaps it almost obvious is to pick an estimation method that will deliver the desired results over a historical testing window. The potential for this is widely discussed in the public discourse on financial regulations and especially whether banks should be allowed to pick the model (the internal ratings approach) or if the authorities should have partial or full control over which model is used. However, there is no guarantee that an estimation method that delivers a favorable result today will deliver similar favorable results in the future, and anyone looking to take excessive risks without being detected is likely to look elsewhere.

A more direct and promising way to take excessive risk, while remaining firmly compliant, is to to cherry pick trades that result in a portfolio with the highest UR. The discussion in this Section is a clear example of this. Even if the testing procedure signals full compliance, the bank still can report risk that is perhaps 10% lower than it actually is. Even more worryingly the very fact that the control mechanism incentivizes the risk taker to load up on tail risk, perversely increasing the banks' default probabilities and even systemic risk.

There are some steps that could be taken to reduce the potential for excessive risk taking. The obvious solution would be to increase the size of the testing window, but that raises issue of having to wait too long for the outcome of the test. One can also make use of in–sample backtests, which of course can be made to use much longer testing windows, but the risk here is that the model designer will know what happened and can therefore tailor the test and portfolio composition to deliver the most desirable outcome.

One could also maintain strict Chinese walls between the risk manager and the risk taker. The less the risk taker knows about the control process the harder it is for her to react ex ante. This is straightforward to do internally in financial institutions and is usually directly embedded in their internal processes. This is not possible to do for financial regulations, since the model for market risk capital is transparent. This is one reason why it is not advisable to require financial institutions to use the same model for regulatory and internal risk control purposes, as many regulators desire. Another alternative would be to use a different risk measure, such as the ES. Nevertheless, as we have shown, it is estimated with less precision.

Besides exploiting the weakness in the control process, an alternative device for excessive risk taking is risk manipulation. Two potential mechanisms that may mitigate risk manipulation have been studied in the literature from a more theoretical point of view. Colliard (2014) argues that internal models incentivize banks to choose model strategically, proposing a remedy based on penalizing banks with ex-ante low risk weights who then suffer what he calls abnormal losses or failure. Another alternative would be to follow the suggestion of Blum (2008) who propose the use of a non-risk sensitive leverage ratio in cases where the supervisors may not be able to identify dishonest bank behavior.

4 Probability shifting

The discussion in Section 2 above indicates that the estimation accuracy of risk measures increases with lower probabilities. Theoretically, the asymptotic variance is fractional to $1/k$, where $1-k/N$ is the probability level in the risk measure. If the probability level falls, k increases and thus the estimation uncertainty falls. This is also validated by the results in in Tables 1 and 2 where the confidence bounds narrow sharply as the probabilities become less extreme. This result suggests that one would be better off estimating a model at a low probability, and scale it up to the probability needed — probability shifting. Such an approach also has the potential to remove the room for underreporting risk, and hence reduce the potential for excessive risk taking.

4.1 Room for underreporting

If one applies the Basel traffic light approach to VaR 90%, the green light zone will allow no more than $l = 30$ violations, where the corresponding exact probability of being in the green zone is then 87.53% (close to but below 90% as in the design of the traffic light system). Suppose we extend the analysis in Section 3 to this case, where we have a bank that either intends to follow the letter of the rule or violate it in specific ways.

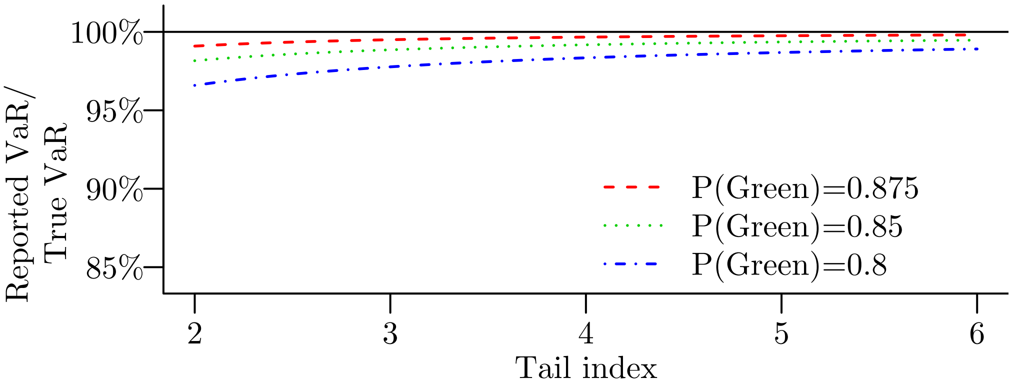
We obtain the UR from (1) with $W = 250$, $l = 30$ and $p = 0.90$, and show the result in Figure 5. Contrasting this figure with Figure 1, two key results emerge. First, the room for underreporting risk is substantially reduced. Even if the bank chooses not to comply, the UR remains small. Second, there is virtually no difference in the UR across different tail thickness, effectively eliminating the preference for heavy tailed assets. The same result holds if one focuses on the red zone.

4.2 Probability shifting

Switching to VaR 90% not only reduces the absolute UR but also reduces banks' incentive for taking on more tail risk, ex ante. However, there are two reasons why this may not be implemented in practice. First, as in the discussion between VaR and ES, VaR can be manipulated, particularly for a VaR at a less extreme probability level. Second, the scaling factor varies across different distributions. Consequently in order to to scale VaR 90% to VaR 99%, it is necessary to pre–estimate the tail thickness. The first issue

Figure 5: Underreporting room with green light targets: VaR 90%

Note: The figure shows the UR ratio against the tail index (α) for three targeted levels of probability falling in the green zone (τ). Here the parameters are $W = 250$, $l = 30$ and $p = 0.90$. The dash (red), dotted (green) and dash-dotted (blue) lines correspond to τ levels at 87.53%, 85% and 80% respectively.



can be solved by moving to ES 90%. As it turns out, this also provides a good solution to the second issue.

By applying the scaling rule from Daníelsson et al. (1998) and Daníelsson et al. (2006), we get that

$$\mathrm{VaR}(99\%) \approx 10^{1/\alpha}\,\mathrm{VaR}(90\%) \approx 10^{1/\alpha}\frac{\alpha-1}{\alpha}\,\mathrm{ES}(90\%).$$

In other words, to scale from ES(90%) to VaR(99%) we need a scaling factor

$$h(\alpha) = 10^{1/\alpha}\frac{\alpha-1}{\alpha}.$$

The first part of the scaling law comes from 90% to 99%, and the second from ES to VaR.

As a practical matter it would be straightforward to estimate α, for example with the Hill (1975) estimator. However, estimating α potentially introduces an additional estimation uncertainty that may distort the precision gain from moving to a lower probability level. Therefore, we tend to choose an h that can comply with most portfolios of banks. Notice that from the Feller theorem (Theorem VIII.8, Feller (1971)), the heaviest tail in a portfolio dominates. The choice of h should depends on a conservative estimate of α, i.e. as low as possible. Empirically, most stock returns have a tail index between 2.5 and 5 which would give $h(2.5) = 1.507$ and $h(5) = 1.268$. Therefore, for market risk analysis, it is proper to choose $h = 1.5$. We use a

Table 3: VaR 99% from probability shifting

Note: the table shows the finite sample performance of the VaR 99% estimates in a similar simulation setup as that for Table 2. The VaR 99% in right column is estimated by first estimating a ES 90% and then scaling that by 1.5.

N	alpha	VaR 99%			VaR 99% from ES 90%		
		bias	se	99% conf	bias	se	99% conf
300 days	2.5	1.11	0.33	[0.61,2.46]	0.93	0.19	[0.63,1.65]
4 years	2.5	1.03	0.15	[0.74,1.51]	0.93	0.10	[0.75,1.30]
10 years	2.5	1.01	0.09	[0.82,1.28]	0.94	0.06	[0.81,1.15]
300 days	3	1.09	0.27	[0.64,2.16]	0.96	0.14	[0.69,1.46]
4 years	3	1.03	0.12	[0.77,1.43]	0.96	0.08	[0.80,1.21]
10 years	3	1.01	0.08	[0.84,1.23]	0.96	0.05	[0.85,1.11]
300 days	4	1.07	0.21	[0.69,1.85]	1.00	0.11	[0.76,1.35]
4 years	4	1.02	0.10	[0.80,1.34]	1.00	0.06	[0.86,1.18]
10 years	4	1.01	0.06	[0.86,1.19]	1.00	0.04	[0.91,1.11]
300 days	5	1.06	0.18	[0.72,1.70]	1.02	0.10	[0.80,1.32]
4 years	5	1.02	0.09	[0.82,1.29]	1.02	0.05	[0.89,1.18]
10 years	5	1.01	0.06	[0.88,1.16]	1.03	0.03	[0.94,1.12]

simulation to demonstrate the accuracy of estimating VaR 99% by ES 90% multiplying 1.5.

The simulation results in Table 3 show that the estimation error falls sharply when using the scaling estimator. This becomes especially notable in overestimation, which is much less of a problem when using the scaling estimator. The underestimation problem is also weakened in most of the cases, with the only exception on the most heavy-tailed case ($\alpha = 2.5$) with the largest sample size (10 years). The downside of the probability shifting approach is that as the scaling constant becomes increasingly inaccurate: compared to the original estimated VaR 99%, the bias of the new estimate increases in some cases.

4.3 Implications

The small sample sizes available for the estimation of risk, typically a few hundred or at most a low thousands, create considerable problems in terms of estimation accuracy and the room to underreport risk. These problems become much smaller as the probabilities fall, and by $p = 90\%$, the estimation results become quite reliable. While, that is not extreme enough for most end-users, it is quite straightforward to estimate the risk at the 90% level and scale to the 99% level, either using a typical scaling coefficient, like the 1.5 above or estimating it. What the results show that the estimation error falls sharply when using the scaling estimator, regardless of whether the scale is exactly correct or not, whereas the bias may increase. Ultimately, this gives the end-user a clear trade-off between accuracy and uncertainty of the estimate.

5 Conclusion

The focus of this paper is on three key issues in risk measuring: accuracy, scope for underreporting and how estimation can be improved. We compare the two most commonly used risk measures, VaR and ES, and find that both are estimated very imprecisely. One needs half a century of daily data for the estimators to reach their asymptotic properties and at sample sizes of few hundred, the risk forecast retain very little information content. From an estimation point of view, VaR is more accurate than ES, but one might still prefer to use ES as it is harder to to manipulate than VaR.

Given existing testing methodologies, there is considerable scope for underreporting VaR without affecting the backtesting results. Particularly worrying is the result that the testing methodologies incentivize financial institutions to increase tail risk.

There are some steps that can be taken to increase the estimation quality and the scope for underreporting, with perhaps the most fruitful, the estimation of risk at a low probability, like 90% and scaling that to the desired probability, perhaps 99%.

We suspect that these general results will not come as a surprise to most empirical researchers. However, the degree of uncertainty was much higher than we have anticipated, and we did not expect the existing literature to be silent on this issue. These results will be more surprising to those users without direct empirical experience with statistical risk measures, like most senior decision-makers in financial institutions and regulatory agencies. Our impression is that statistical risk measures are increasingly being imposed as control tools throughout the financial sector, without adequate understanding of the key empirical issues.

These results have particular implications to how one should use statistical risk measures in decision–making. At the risk of overgeneralizing, there are three different approaches one can take. The first is to treat one carefully-chosen risk forecast as an approximate truth, and use it to allocate portfolios and set risk limits without too many questions asked. This is certainly the view of the financial regulators and many a risk controller and internal auditor. The second approach, which we have often witnessed in our discussion with practitioners, is to use a statistical risk measure as an indication of the underlying risk, one indication amongst several. The risk manager might run several different models in parallel and make use of non-model information in controlling risk. The final approach is to reject statistical risk forecasting altogether. Our results validate the second approach, risk forecasts provide a noisy signal, and so long as the end-user understands the nature of the noise, she is well-placed to use risk forecasts in the best possible way.

While somewhat pessimistic, our conclusion is not that one should not report risk estimates if we only have short samples. Not only is not unfeasible since risk estimates are required in many situations, but also because it is possible to make the risk management process more accurate and have more integrity.

There are several steps users can take to increase the likelihood of success. First, is not to attempt the impossible. If an estimation window of 250 or 500 days is not sufficient to deliver risk estimates with any reasonable accuracy, don't try to do so. Second, explicitly quantify the accuracy of the risk estimate, perhaps by requiring 99% empirical competence bounds. The standard practice in statistics is to require a formal analysis of statistical accuracy. Academic journals require this, the authorities and practitioners generally do not. Having done so, the users could then specify what they see as the minimum acceptable precision in risk estimation and taylor the probability and sample size to that. Third, there are ways one take to improve the estimation, for example by estimating a model at low probability level and applying a probability shift to obtain a more extreme level. Fourth, the users should not limit themselves a single methodological approach, instead implement a range of techniques, including measuring both VaR and ES, and then across multiple probabilities. Finally, when it comes to choosing risk measurement methodologies, don't simply focus on the theoretic properties of the various methodologies under consideration, also study the empirical properties, especially at typical sample sizes.

The main negative conclusion relates to financial regulations because there the inflexibility and long life span of methodologies, coupled with transparency, makes the risk measurement process inaccurate and prone to underreporting. Internally, if allowed to do so by the authorities, financial institutions can avoid many of these problems. This leaves the question of whether given uncertainty of the estimates, the regulatory framework can actually discourage risk–taking? Our conclusion is a qualified yes, but not with current methodologies.

This is a concern because while at the moment, market risk regulations only apply to capital calculations, the authorities have signaled their strong preference for the same method being used for all internal risk calculations, including internal risk management and annual reports. At the same time, other types of financial institutions are increasingly being required to use similar methodologies, European insurance companies now calculate risk by ES and if the current trends towards moving asset managers under a Basel type regulatory umbrella come to fruition, we could also see asset managers being required to use the statistical methods discussed above.

A Appendix

A.1 The number of simulations

The simulations are used not only to obtain estimates of the risk measures, but more importantly the uncertainty of those estimates. This means that in practice we aim to capture the quantiles of the quantiles. Our somewhat ad hock criteria for the results is that they are accurate for at least three significant digits, and as it turns out it requires at least $S = 10^7$ simulations. For the largest sample sizes, we are then generating $S \times N = 10^7 \times 2.5 \times 10^5 = 2.5 \times 10^{12}$ random numbers, and for each sequence need to find a quantile and a mean.

Why is such a large simulation necessary? Taking the VaR measure as an example, from each sample, we obtain one simulated quantity $\hat{q}_F/q_F - 1$. Across S simulated samples, we obtain S such ratios denoted as $r_1, r_2, \cdots, r_S$. They are regarded as i.i.d. observations from the distribution of $\hat{q}_F/q_F$, denoted as F_R. Since we intend to obtain the 99% confidence interval of this ratio, $[F_R^{-1}(0.005), F_R^{-1}(0.995)]$, we take the $[0.005S]$-th and $[0.995S]$-th order statistics among $r_1, \cdots, r_S$, $r_{S,[0.0005S]}$ and $r_{S,[0.995S]}$ to be the estimates of the lower and upper bounds respectively. For the lower bound, following Theorem 2 in Mosteller (1946), we get that as $S \to \infty$,

$$\sqrt{S}\left(\frac{r_{S,[0.0005S]}}{F_R^{-1}(0.005)} - 1\right) \overset{d}{\to} N\left(0, \frac{0.0005 \cdot (1 - 0.0005)}{\left(F_R^{-1}(0.005)\right)^2 f_R^2(F_R^{-1}(0.005))}\right),$$

where f_R is the density function of F_R. Following Proposition 1, the distribution F_R can be approximated by a normal distribution with a given standard deviation σ_N. Using this approximated distribution, we can explicitly calculate the asymptotic variance above as

$$\sigma_R^2 = \frac{0.005 \cdot (1 - 0.005)}{(\sigma_N \Phi^{-1}(0.005))^2 \left(\frac{1}{\sigma_N} \phi \left(\frac{\sigma_N \Phi^{-1}(0.005)}{\sigma_N} \right) \right)^2} = 3.586.$$

Note that this variance is independent of σ_N. Therefore this result can be applied to any estimator that possesses asymptotic normality.

To ensure that the relative error between our simulated lower bound $r_{S,[0.0005S]}$ and the actual lower bound $F_R^{-1}(0.005)$ is less than 0.001 with a confidence level of 95%, the restriction requires a minimum S such that

$$S \geq \sigma_R^2 * \left(\frac{\Phi^{-1}(0.975)}{0.001} \right)^2 = 1.378 \times 10^7.$$

A minimum of $S = 2 \times 10^7$ samples is necessary for our simulation study and that is the number of simulated samples we use throughout this section.

A.2 proofs

Proof of Proposition 1 and 2.
Under the conditions in the proposition, Theorem 2.4.8 in de Haan and Ferreira (2006) showed that there exists a proper probability space with Brownian motions $\{W_N(s)\}_{s \geq 0}$ such that as $N \to \infty$,

$$\left| \sqrt{k} \left(\frac{X_{N,N-[ks]}}{U(N/k)} - s^{-1/\alpha} \right) - \frac{1}{\alpha} s^{-\frac{1}{\alpha}-1} W_N(s) - \sqrt{k} A(N/k) s^{-\frac{1}{\alpha}} \frac{s^{-\rho} - 1}{\rho} \right| \xrightarrow{P} 0 \tag{2}$$

holds uniformly for all $0 < s \leq 1$. By taking $s = 1$, Proposition 1 follows immediately.

To prove Proposition 2, we apply the integral for $s \in (0, 1]$ to (2) and obtain that as $N \to \infty$

$$\sqrt{k} \left(\frac{\hat{e}_F(1 - k/N)}{U(N/k)} - \frac{1}{1 - 1/\alpha} \right) - \int_0^1 \frac{1}{\alpha} s^{-\frac{1}{\alpha}-1} W_N(s) ds - \lambda \frac{1}{(1 - \rho)(1 - 1/\alpha - \rho)} \xrightarrow{P} 0.$$

Notice that it is necessary to have $\alpha > 2$ to guarantee the integrability of $\int_0^1 \frac{1}{\alpha} s^{-\frac{1}{\alpha}-1} W_N(s) ds$.

Similarly, from the inequality (2.3.23) in de Haan and Ferreira (2006), we get that for any $\varepsilon > 0$, with sufficiently large N,

$$\left|\sqrt{k}\left(\frac{U(N/ks)}{U(N/k)} - s^{-1/\alpha}\right) - \sqrt{k}A(N/k)s^{-\frac{1}{\alpha}}\frac{s^{-\rho}-1}{\rho}\right| \leq \varepsilon\sqrt{k}A(N/k)s^{-1/\alpha-\rho-\varepsilon},$$

holds for all $0 < s \leq 1$. With a small ε such that $1/\alpha + \rho + \varepsilon < 1$, we can take integral for $s \in (0,1]$ on both sides and obtain that as $N \to \infty$,

$$\sqrt{k}\left(\frac{e_F(1-k/N)}{U(N/k)} - \frac{1}{1-1/\alpha}\right) \to \lambda\frac{1}{(1-\rho)(1-1/\alpha-\rho)}.$$

Therefore, by comparing the asymptotics of $\frac{\hat{e}_F(1-k/N)}{U(N/k)}$ and $\frac{e_F(1-k/N)}{U(N/k)}$, we get that

$$\sqrt{k}\left(\frac{\hat{e}_F(1-k/N)}{e_F(1-k/N)} - 1\right) \xrightarrow{d} \frac{\alpha-1}{\alpha^2}\int_0^1 s^{-\frac{1}{\alpha}-1}W(s)ds.$$

The proof is finished by verifying the variance of the limit distribution as follows.

$$\begin{aligned}\mathrm{Var}\left(\frac{\alpha-1}{\alpha^2}\int_0^1 s^{-\frac{1}{\alpha}-1}W(s)ds\right) &= \frac{(\alpha-1)^2}{\alpha^4}\int_0^1 ds\int_0^1 dt\left(s^{-\frac{1}{\alpha}-1}t^{-\frac{1}{\alpha}-1}\min(s,t)\right)\\ &= \frac{2(\alpha-1)^2}{\alpha^4}\int_0^1 dt\left(t^{-\frac{1}{\alpha}-1}\int_0^t s^{-\frac{1}{\alpha}}ds\right)\\ &= \frac{2(\alpha-1)}{\alpha^3}\int_0^1 t^{-\frac{2}{\alpha}}dt\\ &= \frac{2(\alpha-1)}{\alpha^2(\alpha-2)}.\end{aligned}$$

■

References

Acharya, V. V., L. H. Pedersen, T. Philippon, and M. Richardson (2010, May). Measuring systemic risk. Working Paper.

Adrian, T. and M. K. Brunnermeier (2016). Covar. *American Economic Review.*

Agarwal, V. and N. Y. Naik (2004). Risks and portfolio decisions involving hedge funds. *Review of Financial studies 17*(1), 63–98.

Alexander, C. and J. M. Sarabia (2012). Quantile uncertainty and value-at-risk model risk. *Risk Analysis 32*(8), 1293–1308.

Artzner, P., F. Delbaen, J. Eber, and D. Heath (1999). Coherent measure of risk. *Mathematical Finance 9*(3), 203–228.

Aussenegg, W. and T. Miazhynskaia (2006). Uncertainty in value-at-risk estimates under parametric and non-parametric modeling. *Financial Markets and Portfolio Management 20*(3), 243–264.

Basel Committee (1996). *Amendment to the Capital Accord to Incorporate Market Risks.* Basel Committee on Banking Supervision. `http://www.bis.org/publ/bcbs24.pdf`.

Basel Committee on Banking Supervision (2014). Fundamental review of the trading book: outstanding issues. Technical report, Basel Committee on Banking Supervision.

Berkowitz, J. and J. O'Brien (2002). How accurate are value-at-risk models at commercial banks? *Journal of Finance 57*(3), 1093–1111.

Blum, J. M. (2008). Why Basel II may need a leverage ratio restriction. *Journal of Banking & Finance 32*(8), 1699–1707.

Brownlees, C. T. and R. F. Engle (2015). SRISK: A conditional capital shortfall measure of systemic risk.

Christoffersen, P., V. Errunza, K. Jacobs, and H. Langlois (2012). Is the potential for international diversification disappearing? a dynamic copula approach. *Review of Financial Studies 25*(12), 3711–3751.

Colliard, J.-E. (2014). Rational blinders: strategic selection of risk models and bank capital regulation. *ECB Working Paper No. 1641.*

Csörgő, M. and L. Horváth (1993). *Weighted approximations in probability and statistics*. Wiley.

Cummins, J. D., D. Lalonde, and R. D. Phillips (2004). The basis risk of catastrophic loss index securities. *Journal of Financial Economics 71*(1), 77–111.

Cuoco, D. and H. Liu (2006). An analysis of var-based capital requirements. *Journal of Financial Intermediation 15*, 362–394.

Daníelsson, J., C. de Vries, B. Jorgensen, G. Samorodnitsky, and S. Mandira (2012, March). Fat tails, VaR and subadditivity. *Journal of Econometrics*.

Daníelsson, J., P. Hartmann, and C. G. de Vries (1998, June). The cost of conservatism: Extreme returns, value-at-risk, and the Basle ḿultiplication factor: *Risk January 1998*. www.RiskResearch.org.

Daníelsson, J., K. James, M. Valenzuela, and I. Zer (2016). Model risk of risk models. *Journal of Financial Stability*.

Daníelsson, J., B. N. Jorgensen, M. Sarma, and C. G. de Vries (2006). Comparing downside risk measures for heavy tailed distributions. *Economics letters 92*(2), 202–208.

de Haan, L. and A. Ferreira (2006). *Extreme value theory: an introduction*. Springer.

Einmahl, J. (1992). Limit theorems for tail processes with application to intermediate quantile estimation. *Journal of Statistical Planning and Inference 32*(1), 137–145.

Fama, E. (1963). Mandelbrot and the stable Paretian hypothesis. *Journal of Business 36*(4), 420–429.

Feller, W. (1971). *An introduction to probability theory and its applications*, Volume II. New York: Wiley.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association 106*(494), 746–762.

Guptaa, A. and B. Liang (2005). Do hedge funds have enough capital? a value-at-risk approach. *Journal of Financial Economics 77*, 219253.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist. 35*, 1163–1173.

Ibragimov, R., D. Jaffee, and J. Walden (2011). Diversification disasters. *Journal of Financial Economics*.

Mandelbrot, B. B. (1963). The variation of certain speculative prices. *Journal of Business 36*, 392 – 417.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance 7*, 77–91.

Marshall, D. A. and E. S. Prescott (2006). State-contingent bank regulation with unobserved actions and unobserved characteristics. *Journal of Economic Dynamics & Control 30*, 2015–2049.

Mosteller, F. (1946). On some useful "inefficient" statistics. *The Annals of Mathematical Statistics 17*(4), 377–408.

O'Brien, J. and P. J. Szerszen (2014). An evaluation of bank var measures for market risk during and before the financial crisis. working paper, Federal Reserve Board.

Patton, A. J. (2009). Are "market neutral" hedge funds really market neutral? *Review of Financial Studies 22*(7).

Perotti, E., L. Ratnovski, and R. Vlahu (2011). Capital regulation and tail risk. Technical report, IMF.

RiskMetrics Group (1993). *RiskMetrics-technical manual.* J.P. Morgan.