TECHNIQUES FOR VERIFYING THE ACCURACY OF RISK MEASUREMENT MODELS

Paul H. Kupiec

is a senior economist with the Board of Governors of the Federal Reserve System in Washington, D.C.

Risk exposures are typically quantified in terms of a "value at risk" (VaR) estimate. A VaR estimate corresponds to a specific critical value of a portfolio's potential one-day profit and loss probability distribution. Given their function both as internal risk management tools and as potential regulatory measures of risk exposure, it is important to quantify the accuracy of an institution's VaR estimates.

This study shows that the formal statistical procedures that would typically be used in performance-based VaR verification tests require large samples to produce a reliable assessment of a model's accuracy in predicting the size and likelihood of very low probability events. Verification test statistics based on historical trading profits and losses have very poor power in small samples, so it does not appear possible for a bank or its supervisor to verify the accuracy of a VaR estimate unless many years of performance data are available. Historical simulation-based verification test statistics also require long samples to generate accurate results: Estimates of 0.01 critical values exhibit substantial errors even in samples as large as ten years of daily data.

This study considers alternative statistical techniques that could be used to verify the accuracy of estimates of the tail values of the distribution of potential gains and losses for a portfolio of securities, futures, and derivative positions. These so-called reality checks have been advanced as a tool for determining the accuracy of risk exposure estimates generated by risk measurement models.

Dealer banks and broker-dealers typically maintain internal risk measurement models that are used to estimate the daily global exposures generated by the institution's portfolio of financial assets and derivative obligations (see "Derivatives" [1993, Appendix III]). Risk exposures are typically quantified in terms of a "value at risk" (VaR) estimate.

A VaR estimate corresponds to a specific critical value of a portfolio's potential one-day profit and loss probability distribution. Typically, a VaR estimate is defined to be a loss large enough so that the probability that the portfolio could post a larger loss is at most some specified value, like 1% or perhaps 5%. A VaR measure thus corresponds to a specific left-hand critical value of the portfolio's potential profit and loss distribution.

The Basle Bank Supervisors Committee proposes that critical value estimates from a bank's internal risk measurement model become the basis for a bank's market risk regulatory capital requirement. Its proposal defines VaR in terms of a two-week holding period (see "An Internal Model-Based Approach" [1995]). Similarly, under a proposal by the Derivatives

Policy Group (DPG), internal model risk exposure estimates could be used to establish capital guidelines for the derivatives activities of unregulated affiliates of U.S. broker-dealers (see "A Framework for Voluntary Oversight" [1995]).

Given their function both as internal risk management tools and as potential regulatory measures of risk exposure, it is important to quantify the accuracy of an institution's model-based risk exposure measures. Despite the importance of accuracy assessment, little research has considered the statistical techniques that would be appropriate for judging the quality of a financial institution's VaR estimates.[1]

The Group of Thirty "Derivatives" study [1995] suggests that institutions perform "reality checks" for judging model performance. Recommendation 8, for example, suggests that an institution's VaR estimates be compared against its portfolio's subsequent profit and loss outcomes, but it does not provide any detail regarding the formal statistics that facilitate the comparison. Similarly, the Basle Supervisors Committee recommends backtesting as a means of verifying the accuracy of a bank's risk exposure estimates, but again the recommendation does not provide the details of the proposed verification test.

This study derives the formal statistical properties of alternative statistics that can be used to verify the accuracy of a VaR estimate. The procedures use historical profits and losses on the institution's portfolio or historical simulation exercises to verify the accuracy of an institution's estimate of its potential loss exposure.

The results indicate that, unless a relatively long performance history or historical simulation data base is available, there are significant statistical difficulties surrounding verification of VaR estimates. The results have implications both for banks that wish to assess the accuracy of their internal risk measurement models as well as for supervisors who must verify the accuracy of an institution's risk measurement model and assign an appropriate market risk scaling factor under the Basle proposal.

Given the nature of internal model risk exposure estimates, performance-based verification tests should compare one-day potential loss estimates with one-day actual performance data. Although the one-day horizon of comparison may be self-evident to an institution assessing the performance of its own internal risk measurement model, analysis suggests that the one-day horizon is equally appropriate for a supervisor attempting to verify an internal model-based capital charge.

Even when tests are based on daily performance comparisons, small sample test statistics have extremely poor power for detecting a model or institution that habitually underestimates potential loss amounts. If only a small history of performance is available, moreover, a model or institution can substantially underestimate the magnitude of its potential losses with little probability of detection either internally by the bank's risk management staff or externally by a supervisor using a performance-based verification test. Reliable performance-based verification techniques require a relatively long comparison sample period.

Verification schemes need not be based on historical performance. If the distributions of the underlying financial factors are stationary, loss exposure estimates can, in theory, be corroborated using the critical value estimates from simulations of the historical loss distribution of an institution's current portfolio. The results presented here show that historical simulation-based verification schemes also perform very poorly unless historical simulation sample sizes are large.

When potential loss distributions are fat-tailed, simulation-based critical value estimates exhibit significant biases and have standard errors of substantial magnitude, even in relatively large samples. The characteristics of simulation-based verification tests do not recommend their use either as a technique for estimating tail values or as a means of performing validation checks of risk exposure estimates.

Because reliable performance-based verification tests require significant amounts of data, the verification process is necessarily time-consuming. If a model is determined to be inaccurate, the model will be altered, the institution will begin accumulating a new performance data sample, and a substantial amount of time must elapse before the accuracy of the new model can be confidently accepted. Time considerations have implications for the VaR multiplication factor of the Basle proposal. If the magnitude of the factor is linked to statistical verification, in the event a bank's model is deemed inaccurate, the results suggest that the bank's VaR scaling factor should remain elevated for a significant time period.

I. THE REGULATORY VERIFICATION PROBLEM

Under the internal models proposal for setting market risk capital requirements, banks would use their internal risk measurement models to estimate the distribution of potential loss exposure associated with their trading portfolio positions.[2] Regulation would require banks to report an estimate of the size of the potential loss that would be exceeded less than 1% of the time in a two-week period. This loss estimate would, in effect, be an estimate of the 1% left-hand critical value of the trading account's two-week potential profit and loss distribution.

Market risk capital requirements would be some multiple — the "scaling factor" — of the loss associated with the 1% critical value reported by the bank. The scaling factor would (under the current version of the proposal) have a minimum value of 3, which would be increased if the supervisor concludes that a bank's risk measurement model is inaccurate.

Despite the central importance of model verification, there is no commonly accepted standard statistical approach for determining the accuracy of VaR estimates. A typical VaR model estimates potential changes in portfolio value by approximating the value changes for the component instruments as linear functions of the changes in their underlying pricing factors (e.g., default-free interest rates of different maturities). The coefficients of the pricing factors are derived from the coefficients in a Taylor series expansion of a theoretical pricing model (see "RiskMetrics" [1995, pp. 108-131]).

As VaR models are not statistical regression models, there is no ex ante measure of their goodness-of-fit. Presumably a model-based loss estimate could be considered accurate if a bank's actual losses do not frequently exceed its ex ante internal model-based critical value estimates, or if its loss estimates do not exceed the potential losses that would have been generated by the portfolio if it had been held by the bank through some historic period.

The Monitoring Horizon

The use of historical performance data to verify internal model-based estimates of long holding-period potential loss exposures is complicated by the endogenous nature of the portfolio's risk. (For a more detailed discussion, see Kupiec and O'Brien [1995].) Over the regulatory monitoring interval, the bank can and will adjust its trading risk exposure.

Any scheme that attempts to verify a long-horizon risk exposure estimate with actual portfolio profits and losses is comparing the risk estimate for a portfolio of fixed composition to the profit or loss performance generated by a series of portfolios that differ in composition. Indeed, even over a single day, an institution's risk profile can be significantly altered by intraday changes in positions. If intraday exposure changes are significant, verification tests should be based on performance calculated by re-marking the original portfolio to market.

Any true long-horizon risk exposure verification scheme would also have to be based on the long-horizon profits or losses generated by repricing the initial portfolio. Monitoring schemes based on historical performance are internally consistent only when they compare a portfolio's potential loss estimate with the same portfolio's actual performance.

A final consideration is the time period necessary to conduct meaningful verification analysis and detect under-reporting banks. The statistical analysis shows that relatively large sample sizes are necessary if verification tests are to have any power against important alternative hypotheses. The time necessary to accumulate a large enough sample of independent two-week horizons is too long to be useful for supervisory purposes. Thus, both data processing and statistical power considerations suggest one-day performance comparisons.

Alternative Statistical Methods for Regulatory Monitoring

If portfolio performance can be monitored, day-to-day profits and losses determine the outcome of a binomial event: Either the bank's loss on trading activities is less than its ex ante estimate (a success), or the loss on trading activities exceeds the ex ante estimate (a failure). If daily forecasts are efficient, potential loss estimates are independent across days, and the performance data are distributed as a series of draws from an independent Bernoulli distribution. Because a supervisor has no knowledge of the parametric form of the bank's profit and loss distribution — and indeed there is good reason to believe that the form of the distribution changes depending on the composition of the bank's portfolio — the size of the differences between a bank's potential loss estimate and its actual gains or losses is not informative.

The null hypothesis — that the probability of a failure on any day is 1% — can be tested in a variety of ways.[3] The appropriate test depends on how the bank is being monitored and the performance comparison sample size available. If the bank is monitored continuously, and a single failure is observed, the supervisor can formally test the hypothesis that the bank's true failure rate is the 0.01 used for its reported VaR. An alternative approach is to monitor the bank at less frequent intervals, and test the null hypothesis using the proportion of failures observed in the monitoring period.

As an alternative to monitoring a bank's actual performance, the bank's loss estimates can be compared periodically to a simulated performance distribution. In this approach, the critical values of the bank's portfolio loss distribution are generated by historically simulating the day-to-day gains and losses the bank's current portfolio would have generated if it were held over some fixed historical time period. This verification technique is termed the historical simulation approach.

II. THE INTERNAL VERIFICATION PROBLEM

Institutions must also assess the accuracy of their risk measurement models. The task of assessing the accuracy of an institution's VaR estimates is statistically the same, whether the assessment is made by a supervisor or by an institution's risk management staff. Although an institution's internal staff could have better information about the parametric form of the potential profit or loss distribution, in verifying tail loss estimates it appears that such knowledge will have little additional value.

Verification tests can be constructed that use information about the parametric form of a portfolio's potential profit and loss distribution. For example, Crnkovic and Drachman (CD) [1995] use Kuiper's [1962] results to construct a goodness-of-fit measure for an estimate of the entire profit and loss distribution. Using a symmetric weighting function, CD specialize their goodness-of-fit measure into a test of the accuracy of a risk management model's tail probability estimates.

Their weighting function places equal importance on the extreme profit and loss tail events. Using a weighting function that implicitly assumes the underlying distribution is symmetric, CD conclude from a Monte Carlo analysis that their testing procedure requires a minimum of 1,000 observations to be reliable.[4] If the symmetry assumption is discarded, the test would require additional data. Compared to CD's testing procedure, the VaR verification tests proposed in this study are computationally simpler and may be more accurate in smaller data sets.[5]

Consequently, even from the perspective of an institution's internal risk management staff, it is appropriate to construct VaR verification tests from the series of Bernoulli trial outcomes generated by a daily performance comparison. Alternatively, if a performance history is not available, an institution might attempt to verify its VaR estimate by comparing it to the critical value of the portfolio's simulated historical loss distribution.

III. VERIFICATION TESTS BASED ON TIME UNTIL FIRST FAILURE

In a performance-based verification scheme, the initial monitoring statistic of interest is the number of observations until a failure is observed. A subsequent section develops the verification tests required when analyzing a monitoring period that covers multiple failures.

Let $\tilde{T}$ be a random variable that denotes the number of days until the first failure is recorded. If p is the probability of a failure on any given day, the probability of observing the first failure in period V is given by:

$$\mathrm{Prob}(\tilde{T} = V) = p(1 - p)^{V-1} \tag{1}$$

$\tilde{T}$ has a geometric distribution with an expected value — the expected number of observations until the first failure is observed — of (1/p). For example, when p = 0.01, the average time until the first failure is 100; when p = 0.05, the average time until failure is 20.

Given a realization for $\tilde{T}$, we want to test that the underlying potential loss estimates are consistent with the null hypothesis. A hypothesis test can be constructed using the likelihood ratio (LR) test procedure. The Neyman-Pearson lemma establishes that the LR test is uniformly most powerful against simple alternative hypotheses in this context.

Given a value for $\tilde{T}$, $\tilde{T}$ = V, the LR statistic for testing the null hypothesis $p = p^*$ is given by LR (V, p^*):

$$LR(V, p^*) = -2\text{Log}\,[p^*(1-p^*)^{V-1}] + 2\text{Log}\,[(1/V)(1-1/V)^{V-1}] \qquad (2)$$

Under the null hypothesis, LR (V, p^*) has a chi-square distribution with 1 degree of freedom.[6] The 5% critical value of this distribution is 3.841; that is, if the likelihood ratio exceeds 3.841, the null hypothesis that $p = p^*$ can be rejected at a 5% Type I error rate. The Type I error rate is the probability of incorrectly rejecting a true null hypothesis.

Exhibit 1 reports the acceptance regions for the 5% [TUFF (0.05)] and 10% [TUFF (0.10)] levels of the time until first failure (TUFF) test for alternative null hypotheses. Notice that the non-rejection region grows larger as the null hypothesis values of p^* approach zero.

When testing the null hypothesis $p^* = 0.01$, the TUFF (0.05) critical values for V are V = 6 and V = 439. That is, if the first failure occurs before the seventh trading day, it can be concluded that $p > 0.01$. If the first failure occurs after the 438th trading day, it can be concluded that $p < 0.01$.

For a null hypothesis of $p^* = 0.05$, the TUFF (0.05) test critical values for V are: between 0 and 1 (an impossibility), and 87. The lower critical value implies that it is impossible to determine at the 5% level whether $p > 0.05$, because under the null hypothesis a failure occurs with 5% probability on the first draw. If V is greater than 87, it can be concluded at the 5% level that the loss estimates are consistent with a tail loss probability of less than 0.05.

EXHIBIT 1
CRITICAL VALUES FOR THE TUFF TEST

Null Hypothesis Probability p^*	Non-Rejection Region for V 0.05 Type I Error	Non-Rejection Region for V 0.10 Type I Error
0.005	11 < V < 879	21 < V < 729
0.010	6 < V < 439	10 < V < 364
0.015	4 < V < 292	7 < V < 242
0.020	3 < V < 219	5 < V < 182
0.025	2 < V < 175	4 < V < 145
0.030	2 < V < 146	3 < V < 121
0.035	2 < V < 125	3 < V < 103
0.040	1 < V < 109	3 < V < 90
0.045	1 < V < 97	2 < V < 80
0.050	V < 87	2 < V < 72

V is the number of observations until the first failure is recorded.

The test of any null hypothesis for which p^* is 0.05 or larger will be associated with a TUFF (0.05) test non-rejection region that includes samples that realize a failure on the first observation. The implication is that the TUFF (0.05) test will reject these null hypotheses only when the true model error rate is smaller than 0.05. In other words, the TUFF (0.05) test does not have the ability to detect models with failure rates greater than 5%.

If the TUFF test Type I error rate is increased beyond the p^* value associated with the null hypothesis, the test can detect an alternative hypothesis for which $p > p^*$. For example, Exhibit 1 reports the non-rejection region for the TUFF (0.10) test of p^* = 0.05. Notice that the TUFF (0.10) test will reject the null hypothesis if a failure is observed on the first observation.

The TUFF (0.05) test critical values reported in Exhibit 1 indicate very large non-rejection regions for the null hypothesis where p^* is small. It may be somewhat surprising to observe that if the first failure occurs on the seventh trading day, the maximum likelihood estimate of p is (1/7) or 14.3%, and yet a null hypothesis of $p^* = 0.01$ cannot be rejected by the TUFF (0.05) test. Despite the fact that the LR test criterion generates the most powerful test using data

EXHIBIT 2
Selected Type II Error Rates for the TUFF (0.05) Test

Null Hypothesis	Alternative Hypothesis	Type II Error Rate
$p^* = 0.010$	$p = 0.015$	0.898
$p^* = 0.010$	$p = 0.020$	0.868
$p^* = 0.010$	$p = 0.030$	0.808
$p^* = 0.010$	$p = 0.040$	0.751
$p^* = 0.010$	$p = 0.050$	0.698
$p^* = 0.025$	$p = 0.030$	0.908
$p^* = 0.025$	$p = 0.040$	0.884
$p^* = 0.025$	$p = 0.050$	0.857

The Type II error rate is the probability of accepting the false null hypothesis using a 5% level TUFF test when the specific alternative hypothesis is true.

on the time until the first failure, the substantial size of the region over which the null hypothesis cannot be rejected is an indication that the test statistic has a poor ability to distinguish among a wide range of interesting alternative hypotheses.

Exhibit 2 reports the Type II error rates for selected TUFF (0.05) test values. A Type II error is the probability of accepting a false null hypothesis. For example, if the null hypothesis is $p^* = 0.01$, a Type II error is the probability of accepting the null hypothesis $p^* = 0.01$, when in fact the probability of a failure on any single observation is different from 0.01.

The Type II error rate depends on the true underlying value of p that generates the data. The larger the true probability, the smaller the probability of committing a Type II error.

The error rates reported in Exhibit 2 show that there is a very high probability that the null hypothesis $p^* = 0.01$ or $p^* = 0.025$ could be accepted by the TUFF (0.05) test even when a bank's true tail probability is far in excess of the null hypothesis value being tested. The high Type II error probabilities indicate that the TUFF test has very poor power characteristics.[7]

For example, Exhibit 2 shows that if the first failure is observed between the 7th and 438th observation, the null hypothesis $p^* = 0.01$ cannot be rejected by the TUFF (0.05) test. Yet if the portfolio's true probability of experiencing a loss worse than the reported 1% level is $p = 0.02$, 86.8% of the time the TUFF (0.05) test would accept the false null hypothesis.

The difficulty of distinguishing between very small alternative VaR values (e.g., 0.01 and 0.02) may at first glance appear to be a point of academic interest with little practical significance. In fact, such small differences can have substantial economic importance.

Consider a fat-tailed probability distribution often used to model financial asset prices, such as the Student t-distribution. The 0.01 critical value of a t-distribution with 1 degree of freedom is –31.82. The 0.02 critical value from the same distribution is –15.89. In this situation, a 0.02 cumulative probability tail loss underestimates the 99% VaR value by 100%. The particularly troubling aspect is that it would be virtually impossible for the regulatory authority or the bank's internal risk management staff to detect even such a gross underestimate using the TUFF (0.05) test.

Although the 1-degree of freedom t-distribution is (intentionally) a dramatic illustration of the problem, the qualitative point of the example is general: Slight differences in the cumulative probability attached to a VaR estimate can translate into substantial differences in potential loss amounts.

The TUFF (0.05) test Type II error rates can be reduced by accepting a greater probability of incorrectly rejecting a true null hypothesis. Exhibit 3 reports the null hypothesis acceptance regions and the corresponding Type II error rates for the alternative $p = 0.02$ for the TUFF test of $p^* = 0.01$ under alternative Type I error rates. The Type II error rates reported show that the TUFF test of the null hypoth-

EXHIBIT 3
Trade-Off Between Type I and Type II Error Rates for the TUFF Test

Level of Tests (Type I Error)	Acceptance Region	Type II Error Probability When True $p = 0.02$
0.05	$6 < V < 439$	0.868
0.10	$10 < V < 364$	0.784
0.15	$15 < V < 319$	0.722
0.20	$20 < V < 287$	0.651
0.25	$24 < V < 262$	0.598

esis p* = 0.01 has poor power against the alternative p = 0.02 even for large Type I error rates.

The analysis suggests that the TUFF statistic has poor ability to distinguish reliably between alternative underlying values for the tail probability associated with a VaR estimate. Although the TUFF test is logically the first test to employ when initially undertaking a performance-based monitoring scheme, its power is limited by the small sample sizes for which it applies.

IV. PERFORMANCE TESTS BASED ON PROPORTION OF FAILURES

Continued monitoring beyond an observed failure will clearly add information that can be used to verify potential loss estimates. Provided the null hypothesis is not rejected, there are alternative ways to analyze additional performance data.[8]

Tests based only on the time between failures are inherently inefficient because they ignore information about the total number of failures that has occurred since monitoring began. When the TUFF test cannot reject the null hypothesis, verification tests should be based on the proportion of failures in the sample.

The probability of observing x failures regardless of order in a sample of size n is:

$$\text{binomial } [n, x]\ (1 - p)^{n-x}\ p^{x} \qquad (3)$$

where binomial [n, x] signifies the binomial coefficient for n objects taken x at a time, and p is the probability of a failure on any one of the independent trials. The likelihood ratio test of the null hypothesis is again the uniformly most powerful test for a given sample size.

The LR test statistic is given by:

$$-2\text{Log}\ [(1 - p^*)^{n-x}\ (p^*)^{x}] +$$

$$2\text{Log}\ [(1 - [x/n])^{n-x}\ (x/n)^{x}] \qquad (4)$$

where p* is the probability of a failure under the null hypothesis, n is the sample size, and x is the number of failures in the sample. We will call the test given in expression (4) the PF (proportion of failures) test. Under the null hypothesis, p = p*, the PF test has a chi-square distribution with 1 degree of freedom. For the specialized case of a single failure in a sample size of n, the PF test is mathematically identical to the TUFF test. Given this equivalence, this test also has the same poor power properties described at length earlier.

In a daily monitoring scheme, the PF test is used to compare the total number of failures observed to the total accumulated sample size. Exhibit 4 enumerates the critical values of n (the sample size rejection regions) that are associated with alternative values for x (the number of observed failures) for testing alternative null hypotheses using the PF (0.05) test.

EXHIBIT 4

MAXIMUM SAMPLE SIZE (n) FOR WHICH THE NULL HYPOTHESIS p = p* IS REJECTED BY A PF (0.05) TEST

Number of Failures	p* = 0.01	p* = 0.02	p* = 0.03	p* = 0.04	p* = 0.05
x = 1	6	3	—	—	—
x = 2	34	17	11	9	—
x = 3	75	38	26	19	16
x = 4	125	63	42	32	26
x = 5	180	91	61	46	37
x = 6	240	121	81	61	49
x = 7	302	152	102	77	62
x = 8	367	184	124	93	75
x = 9	434	218	146	110	88
x = 10	503	253	169	127	102

For example, if two failures are observed in a sample, if the sample size is less than or equal to thirty-four, the null hypothesis p* = 0.01 can be rejected at the 5% level.

For example, assume that six failures are observed in a monitoring period. The values in Exhibit 4 indicate that the null hypothesis p* = 0.01 can be rejected if there are fewer than 241 days in the monitoring period. In other words, it would require six failures in less than one year to reject p* = 0.01. Using the same six-failure example, the null hypothesis p* = 0.05 would be rejected by the performance data

EXHIBIT 5
NON-REJECTION REGIONS FOR PF (0.05) TEST FOR ALTERNATIVE SAMPLE SIZES

Null Hypothesis Probability p*	Non-Rejection Region for x, n = 255 days	Non-Rejection Region for x, n = 510 days	Non-Rejection Region for x, n = 1,000 days
0.010	x < 7	1 < x < 11	4 < x < 17
0.025	2 < x < 12	6 < x < 21	15 < x < 36
0.050	6 < x < 21	16 < x < 36	37 < x < 65
0.075	11 < x < 28	27 < x < 51	59 < x < 92
0.100	16 < x < 36	38 < x < 65	81 < x < 120

x is the number of failures that could be observed in a sample size equal to the specified number of trading days without rejecting the indicated null hypothesis at the 5% level of the PF test.

if the monitoring period covers fewer than fifty days.

Instead of following the continuous monitoring scheme, it might be more convenient to collect a sample of performance-generated Bernoulli outcomes and perform a verification test for a fixed sample size. Exhibit 5 reports the critical number of failures that correspond to PF (0.05) tests for alternative sample sizes and null hypotheses.

Like the TUFF test, the PF test has poor power characteristics in small samples. That is, in small samples the null hypothesis acceptance regions are large, so there is a significant probability that one will accept the null hypothesis when it is false. Exhibit 6 reports the Type II error rates that correspond to selected PF (0.05) hypothesis testing situations. Notice that large sample sizes are required to reduce the PF (0.05) test Type II error rates.[9]

Because Type I and Type II error rates are inversely related, the power of the PF test could also be improved at the expense of increasing the Type I error rate in a reduced sample size, but the trade-off may not be acceptable in many circumstances. For example, suppose a supervisor wants to design a scheme that requires only one year of data and would detect at least 75% of all banks that attempt to report critical value estimates that are twice the 1% regulatory requirement. Such a scheme would allow a bank to record three or fewer failures in a 255-day trading year and still be deemed to meet the regulatory criterion by the supervisor's verification test.

The Type I error rate associated with this rule is about 75%.[10] This implies that in order to catch 75% of the banks under-reporting their potential loss exposures, the supervisor must be willing to falsely accuse of under-reporting 75% of all banks that are reporting accurately. As this example makes concrete, the trade-off between Type I and Type II errors is not very favorable even in samples as long as a year.

EXHIBIT 6
TYPE II ERROR RATES FOR THE PF (0.05) TEST

Null Hypothesis	Alternative Hypothesis	Type II Error Rate n = 255	Type II Error Rate n = 510	Type II Error Rate n = 1,000
p* = 0.010	p = 0.011	0.976	0.949	0.930
p* = 0.010	p = 0.020	0.749	0.557	0.218
p* = 0.010	p = 0.030	0.355	0.101	0.003
p* = 0.010	p = 0.040	0.113	0.008	0.000
p* = 0.025	p = 0.028	0.920	0.941	0.928
p* = 0.025	p = 0.030	0.898	0.901	0.844
p* = 0.025	p = 0.040	0.674	0.523	0.237
p* = 0.025	p = 0.050	0.374	0.154	0.014
p* = 0.050	p = 0.055	0.944	0.913	0.899
p* = 0.050	p = 0.060	0.905	0.819	0.729
p* = 0.050	p = 0.075	0.639	0.329	0.102
p* = 0.050	p = 0.100	0.147	0.009	0.000
p* = 0.075	p = 0.083	0.915	0.903	0.846
p* = 0.075	p = 0.100	0.669	0.478	0.186

The Type II error rate is the probability of accepting the indicated false null hypothesis when the specific alternative hypothesis is true using the PF test with a sample size = n.

V. VERIFICATION SCHEMES BASED ON HISTORICAL SIMULATION

As an alternative to verification tests based on historical profit and loss performance, it is sometimes suggested that historical simulations can be used as a validation technique. Given a portfolio, it is possible to calculate the daily changes in value the portfolio would have experienced if it had been held over some prior period. The daily changes in portfolio value that would have resulted from the historical day-to-day changes in market prices and interest rates could be used to construct a sample histogram. From such a histogram, 1% (or 5%) critical value loss estimates could be determined. A comparison of a VaR estimate with such a simulation-based critical value loss estimate could be the basis of a verification test.

Such an approach assumes that the statistical processes that generate asset price changes are stationary over time. An appealing quality of this approach is that it does not make explicit assumptions about the underlying covariance structures among asset price changes. The historical volatilities and correlations are automatically captured in the historical simulation exercise.

The drawback of such an approach is the large sampling errors associated with their empirical critical value estimates. Historical simulation-based frequency distributions are estimates of the true underlying distribution and consequently are subject to estimation error. In many cases, very large samples are necessary to reduce the sampling error associated with the critical value estimates for very small (or very large) cumulative probability values.

It is possible to derive a theoretical approximation for the variance of the critical value from a sample frequency distribution. Let X_p correspond to the p% critical value of a probability density f (x) so that the integral from negative infinity to X_p equals p%. It can be shown (see Kendall and Stuart [1960, pp. 236-237]) that the variance of an estimate of X_p from a sample of size n is approximately equal to:

$$\mathrm{Var}\,(X_p) \approx p\,(1 - p)/[n\, f\,(X_p)^2] \tag{5}$$

For example, the standard error of the 0.01 critical value estimate from a sample of size n from a normal distribution with a variance of σ^2 is approximately

$$3.7689\ \sigma\ n^{-1/2}$$

The standard error of the 0.05 critical value from this sample is approximately

$$2.1304\ \sigma\ n^{-1/2}$$

These examples illustrate the general property that the standard errors of critical value estimates from sample histograms increase as the cumulative tail probability associated with the critical value declines.

The importance of sampling error in historical simulation-derived estimates of critical values can be illustrated more concretely using Monte Carlo experiments. In our experiments, 10,000 independent samples of various sizes are drawn from known underlying probability distributions. For each of the 10,000 samples, a sample histogram is constructed, and important critical values — the 0.01, 0.05, and 0.10 cumulative probability critical values — are estimated. From these estimates, sampling distributions are constructed for the alternative critical value estimates. The process is repeated for alternative underlying distributions.

Exhibit 7 reports the simulation results for the standard normal distributions for sample sizes of 100, 250, 500, 1,000, and 2,500 observations. Exhibit 8 reports parallel results for the Student-t distribution with 8 degrees of freedom. Exhibit 9 reports results for a t-distribution with 2 degrees of freedom. The progression of distributions from Exhibit 7 through Exhibit 9 includes distributions with increasingly large tail probability weights in their theoretical density functions.

The empirical results reported in Exhibits 7 through 9 illuminate some clear patterns of interest. As the sample size increases, on average, the bias in critical value estimates decreases. The results also show clearly that the standard error of a critical value estimate increases as the cumulative probability associated with the critical value estimate decreases.

Consistent with the theoretical approximation, the standard errors of critical value estimates decline as the sample sizes used to generate the critical value estimates are increased. Similarly, as the sample size increases, the range of critical value estimates record-

EXHIBIT 7
ACCURACY OF HISTORICAL SIMULATION-BASED CRITICAL VALUE ESTIMATES FOR A STANDARD NORMAL DISTRIBUTION FOR A SAMPLE OF SIZE n

Statistic	Theoretical Value	n = 100	n = 250	n = 500	n = 1,000	n = 2,500
0.01 Critical Value	–2.326	–2.148	–2.256	–2.285	–2.307	–2.317
Standard Deviation		0.309	0.209	0.159	0.116	0.074
Minimum Value		–3.658	–3.157	–2.910	–2.828	–2.595
Maximum Value		–1.212	–1.393	–1.755	–1.867	–1.987
0.05 Critical Value	–1.645	–1.594	–1.624	–1.634	–1.638	–1.643
Standard Deviation		0.203	0.130	0.094	0.066	0.042
Minimum Value		–2.469	–2.136	–2.008	–1.910	–1.808
Maximum Value		–0.899	–1.119	–1.289	–1.401	–1.489
0.10 Critical Value	–1.282	–1.254	–1.271	–1.275	–1.278	–1.280
Standard Deviation		0.177	0.107	0.077	0.053	0.034
Minimum Value		–1.865	–1.727	–1.579	–1.476	–1.440
Maximum Value		–0.705	–0.811	–0.989	–1.082	–1.140

The estimates are based on the results from 10,000 sample histograms simulated in S-Plus.

EXHIBIT 8
ACCURACY OF HISTORICAL SIMULATION-BASED CRITICAL VALUE ESTIMATES FOR A t-DISTRIBUTION WITH 8 DEGREES OF FREEDOM FOR A SAMPLE OF SIZE n

Statistic	Theoretical Value	n = 100	n = 250	n = 500	n = 1,000	n = 2,500
0.01 Critical Value	–2.896	–2.636	–2.787	–2.834	–2.867	–2.884
Standard Deviation		0.528	0.366	0.276	0.203	0.128
Minimum Value		–6.898	–4.749	–4.229	–3.832	–3.469
Maximum Value		–1.338	–1.794	–1.814	–2.279	–2.390
0.05 Critical Value	–1.859	–1.803	–1.839	–1.848	–1.855	–1.857
Standard Deviation		0.270	0.176	0.125	0.090	0.057
Minimum Value		–3.111	–2.649	–2.367	–2.303	–2.092
Maximum Value		–1.000	–1.178	–1.333	–1.550	–1.636
0.10 Critical Value	–1.397	–1.372	–1.387	–1.391	–1.395	–1.395
Standard Deviation		0.202	0.132	0.092	0.065	0.041
Minimum Value		–2.237	–1.926	–1.735	–1.709	–1.566
Maximum Value		–0.680	–0.882	–1.046	–1.169	–1.252

The estimates are based on the results from 10,000 sample histograms simulated in S-Plus.

ed for any critical value level declines.

A very important pattern is visible in the results for the t-distribution: 1) the bias of 0.01 and 0.05 critical value estimates increases as the underlying distribution becomes more leptokurtotic (fat-tailed); and 2) the standard error of the 0.01 critical value estimates (and to a lesser degree, the 0.05 critical value estimates) increases markedly as the underlying distribution becomes more leptokurtotic. The 0.01 critical value estimates for the t-distribution with 2 degrees of freedom exhibit strong bias and standard errors that are very large (relative to underlying theoretical critical values) even in sample sizes as large as 2,500 — a sample size equivalent to almost ten years of daily data.

These simulation results suggest that historical simulation-based critical value estimates for 0.01 (and 0.05) cumulative probabilities may suffer from significant biases and are subject to large sampling variation. When the underlying distribution has fat tails, historical simulation-based critical value estimates for the 0.01 level are remarkably unreliable even in large samples.

The bias and variations exhibited by historical simulation-based 0.01 critical value estimates suggest that they are not very reliable estimates of potential losses. As a consequence, they are not suitable benchmarks for comparisons with alternative internal model-generated potential loss estimates. This analysis does not support the use of historical simulations for validation exercises.

VI. CONCLUSION

The statistical results reported here suggest that simple performance-based VaR verification tests require large samples to produce a reliable accuracy assessment. Small sample reality check statistics based on historical trading profits and losses have very poor power against even substantially larger alternative tail probabilities.

The results indicate that there are significant statistical difficulties surrounding the verification of VaR estimates even in performance samples as long as a year. It does not appear possible for a bank or its supervisor to verify the accuracy of a VaR estimate unless a long model performance history is available.

Historical simulation-based reality check statistics also require long historical sample periods to generate accurate results. Historical simulation-based esti-

EXHIBIT 9

ACCURACY OF HISTORICAL SIMULATION-BASED CRITICAL VALUE ESTIMATES FOR A t-DISTRIBUTION WITH 2 DEGREES OF FREEDOM FOR A SAMPLE OF SIZE n

Statistic	Theoretical Value	n = 100	n = 250	n = 500	n = 1,000	n = 2,500
0.01 Critical Value	−6.965	−6.235	−6.725	−6.774	−6.881	−6.933
Standard Deviation		3.357	2.201	1.571	1.142	0.715
Minimum Value		−64.393	−31.651	−22.241	−14.128	−11.064
Maximum Value		−1.601	−2.785	−3.475	−4.242	−4.821
0.05 Critical Value	−2.919	−2.845	−2.887	−2.903	−2.914	−2.918
Standard Deviation		0.738	0.460	0.333	0.233	0.148
Minimum Value		−7.770	−5.193	−4.517	−3.987	−3.613
Maximum Value		−1.153	−1.668	−1.863	−2.190	−2.433
0.10 Critical Value	−1.886	−1.862	−1.874	−1.879	−1.882	−1.885
Standard Deviation		0.390	0.248	0.174	0.121	0.079
Minimum Value		−4.273	−3.016	−2.611	−2.377	−2.165
Maximum Value		−0.786	−1.081	−1.367	−1.425	−1.594

The estimates are based on the results from 10,000 sample histograms simulated in S-Plus.

mates of 0.01 critical values exhibit substantial bias and retain substantial sampling errors even in samples as large as ten years of daily data.

ENDNOTES

The conclusions in this article are those of the author and do not represent the views of the Federal Reserve Board, or any of the Federal Reserve Banks. The author is grateful to Cedomir Crnkovic, Greg Duffee, Mark Fisher, Bob Litterman, Jim O'Brien, Pat Parkinson, Matt Pritsker, and Larry Wall for useful discussions or comments on an earlier draft of this article. Email at pkupiec@frb.gov.

[1] "RiskMetrics—Technical Document" [1995] provides an abbreviated discussion of model verification issues. Crnkovic and Drachman [1995] propose a measure of the goodness-of-fit for an entire estimated potential profit and loss distribution.

[2] As the DPG recommendations do not include a mandatory regulatory capital requirement for the unregulated affiliates of SEC-regulated broker-dealers, our discussion focuses on verification problems for bank regulators.

[3] The null hypothesis of independence could also be tested, but we do not perform such tests in this article.

[4] Such a weighting function may not be appropriate. In historical financial data, the occurrence of extreme losses is more common than the occurrence of an extreme gain. This underlying assymetry will be compounded by a concentration of option positions in an institution's portfolio. See Kupiec and O'Brien [1995] for further discussion.

[5] A formal comparison of these testing techniques is a topic for future research.

[6] The chi-square distribution result is true asymptotically. Monte Carlo simulation results indicate that the critical values of the chi-square distribution provide good approximations in this setting.

[7] The power of a statistical test is defined to be 1 minus the Type II error probability for a given alternative hypothesis.

[8] If the null hypothesis is rejected, and the risk measurement system is altered to correct the detected inaccuracy, performance data from the new model constitute a new data series distinct from the performance data produced by the prior model. The verification process will begin again, and a new set of performance data will accumulate with time.

[9] For each null analyzed, the first reported Type II error rate is calculated for a relatively local alternative hypothesis: an alternative that is 110% of the null. This comparison is included because a 0.01 difference between the null and the alternative is a relatively small difference if the null is 0.1; it is a very large difference if the null is 0.01. The results show that, for all sample sizes, the power against local alternatives declines as the p-value under the null hypothesis shrinks toward 0.

[10] The Type I error rate must be calculated by evaluating the LR test and calculating the corresponding chi-square probability.

REFERENCES

Crnkovic, C., and J. Drachman. "A Universal Tool to Discriminate Among Risk Measurement Techniques." Corporate Risk Management Group, J.P. Morgan & Co., Inc., New York, September 26, 1995.

"Derivatives: Practices and Principles." Washington, D.C.: Group of Thirty, July 1993.

"A Framework for Voluntary Oversight." New York: Derivatives Policy Group, March 1995.

"An Internal Model-Based Approach to Market Risk Capital Requirements." Basle, Switzerland: Basle Committee on Banking Supervision, April 1995.

Kendall, M.G., and A. Stuart. *The Advanced Theory of Statistics*. London: Charles Griffin & Co., 1960.

Kuiper, N.H. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*. Ser. A., Vol. 63 (1962), pp. 38-47.

Kupiec, P., and J. O'Brien. "The Use of Bank Trading Risk Measurement Models for Regulatory Capital Purposes." FEDS Working Paper No. 95-11, Federal Reserve Board, 1995.

"RiskMetrics—Technical Document," third edition. New York: Morgan Guaranty Trust Company Global Research, May 1995.