

**Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.**

A Review of Backtesting and Backtesting Procedures

Sean D. Campbell

2005-21

NOTE: Staff working papers in the Finance and Economics Discussion Series (FEDS) are preliminary materials circulated to stimulate discussion and critical comment. The analysis and conclusions set forth are those of the authors and do not indicate concurrence by other members of the research staff or the Board of Governors. References in publications to the Finance and Economics Discussion Series (other than acknowledgement) should be cleared with the author(s) to protect the tentative character of these papers.

A Review of Backtesting and Backtesting Procedures

Sean D. Campbell*

April 20, 2005

Abstract

This paper reviews a variety of backtests that examine the adequacy of Value-at-Risk (VaR) measures. These backtesting procedures are reviewed from both a statistical and risk management perspective. The properties of unconditional coverage and independence are defined and their relation to backtesting procedures is discussed. Backtests are then classified by whether they examine the unconditional coverage property, independence property, or both properties of a VaR measure. Backtests that examine the accuracy of a VaR model at several quantiles, rather than a single quantile, are also outlined and discussed. The statistical power properties of these tests are examined in a simulation experiment. Finally, backtests that are specified in terms of a pre-specified loss function are reviewed and their use in VaR validation is discussed.

*Board of Governors of the Federal Reserve System. I would like to thank Mike Gibson, Jim O'Brien, Pat Parkinson and Hao Zhou for useful discussions and comments on an earlier draft. The usual disclaimer applies. The views expressed in this paper are those of the author but not necessarily those of the Board of Governors of the Federal Reserve System, or other members of its staff. Address correspondence to Sean Campbell, The Federal Reserve Board, Mail Stop 91, Washington DC 20551.

1 The Current Regulatory Framework

Since 1998, regulatory guidelines have required banks with substantial trading activity to set aside capital to insure against extreme portfolio losses. The size of the set-aside, or market risk capital requirement, is directly related to a measure of portfolio risk. Currently, portfolio risk is measured in terms of its “value-at-risk”. The value-at-risk or VaR of a portfolio is defined to be the dollar loss that is expected to be exceeded only $\alpha \times 100\%$ of the time over a fixed time interval. If, for example, a financial institution reports a 1% VaR of \$10,000,000 over a 1 day horizon then this means that 1% of the time the institution would be expected to realize a loss in excess of \$10,000,000. The current regulatory framework requires that financial institutions use their own internal risk models to calculate and report their 1% value-at-risk, $VaR(0.01)$, over a 10 day horizon.

Market risk capital requirements are directly linked to both the estimated level of portfolio risk as well as the VaR model’s performance on backtests. Specifically, the risk based capital requirement is set as the larger of either the bank’s current assessment of the 1% VaR over the next 10 trading days or a multiple of the bank’s average reported 1% VaR over the previous 60 trading days plus an additional amount that reflects the underlying credit risk of the bank’s portfolio. In precise terms the market risk capital is defined as,

$$MRC_t = \max \left(VaR_t(0.01), S_t \frac{1}{60} \sum_{i=0}^{59} VaR_{t-i}(0.01) \right) + c \quad (1)$$

where S_t reflects the multiplication factor that is applied to the average of previously reported VaR estimates. When a VaR model indicates more risk, the risk based capital requirement rises. What may be less clear from the above expression is that the risk based capital requirement also depends on the accuracy of the VaR model.

Importantly, the multiplication factor, S_t , varies with backtesting results. S_t is determined by classifying the number of 1% VaR violations in the previous 250 trading days, N , into three distinct categories as follows,

$$S_t = \begin{cases} 3.0 & \text{if } N \leq 4 & \text{green} \\ 3 + 0.2(N - 4) & \text{if } 5 \leq N \leq 9 & \text{yellow} \\ 4.0 & \text{if } 10 < N & \text{red} \end{cases} \quad (2)$$

so that a VaR measure which is violated more frequently results in a larger multiplication factor and accordingly a larger risk based capital requirement. Specifically, as long as the 1% VaR has been violated four times or less in the previous 250 trading days or 2 % of the time the multiplication factor remains at its minimal value of three. As the number of violations,

N , increases beyond four so too does the multiplication factor that determines the market risk capital. In the event that more than ten violations of the 1% VaR are recorded in a 250 day span, corresponding to 4% of the sample period, the VaR model is deemed inaccurate and immediate steps are required to improve the underlying risk management system.

This “traffic light” approach to backtesting represents the only assessment of VaR accuracy prescribed in the current regulatory framework. This paper will survey backtesting methods and will discuss how these backtesting methods might improve on the current regulatory framework.

2 The Statistical Approach to Backtesting

2.1 The Definition of Value-at-Risk, VaR

Before describing VaR backtests in any detail it is important to have a concrete definition of VaR in mind. In mathematical terms, a portfolio’s value-at-risk is defined to be the α quantile of the portfolio’s profit and loss distribution,

$$VaR_t(\alpha) = -F^{-1}(\alpha|\Omega_t) \tag{3}$$

where $F^{-1}(\cdot|\Omega_t)$ refers to the quantile function of the profit and loss distribution which varies over time as market conditions and the portfolio’s composition, as embodied in Ω_t , change. The negative sign is a normalization that quotes VaR in terms of positive dollar amounts, i.e. “losses”. As previously mentioned, if today’s 1% VaR , $VaR_t(0.01)$, is \$10,000,000 then this amounts to saying that 1% of the time we should expect to observe a loss on this portfolio in excess of \$10,000,000. Since financial institutions adhere to risk-based capital requirements by using their own internal risk models to determine their 1% value-at-risk, $VaR_t(0.01)$, it is important to have a means of examining whether or not reported VaR represents an accurate measure of a financial institution’s actual level of risk.¹

2.2 The Statistical Framework of VaR Backtests

Since the late 1990’s a variety of tests have been proposed that can be used to gauge the accuracy of a putative VaR model. While these tests differ in their details many of them focus on a particular transformation of the reported VaR and realized profit or loss. Specifically,

¹In general, an institution’s VaR refers to a historical sequence of VaR measurements at a given level of α , $VaR_t(\alpha)$, $t = 1, 2, \dots, T$. To economize on notation we will refer to this object simply as VaR .

consider the event that the loss on a portfolio exceeds its reported VaR , $VaR_t(\alpha)$. Denoting the profit or loss on the portfolio over a fixed time interval, i.e. daily, as $x_{t,t+1}$ then define the “hit” function as follows,

$$I_{t+1}(\alpha) = \begin{cases} 1 & \text{if } x_{t,t+1} \leq -VaR_t(\alpha) \\ 0 & \text{if } x_{t,t+1} > -VaR_t(\alpha) \end{cases} \quad (4)$$

so that the hit function sequence, e.g. $(0, 0, 1, 0, 0, \dots, 1)$, tallies the history of whether or not a loss in excess of the reported VaR has been realized.

Christoffersen (1998) points out that the problem of determining the accuracy of a VaR model can be reduced to the problem of determining whether the hit sequence, $[I_{t+1}(\alpha)]_{t=1}^{t=T}$, satisfies two properties.

1. Unconditional Coverage Property - The probability of realizing a loss in excess of the reported VaR , $VaR_t(\alpha)$, must be precisely $\alpha \times 100\%$ or in terms of the previous notation, $\Pr(I_{t+1}(\alpha) = 1) = \alpha$. If it is the case that losses in excess of the reported VaR occur more frequently than $\alpha \times 100\%$ of the time then this would suggest that the reported VaR measure systematically understates the portfolio’s actual level of risk. The opposite finding of too few VaR violations would alternatively signal an overly conservative VaR measure.
2. Independence Property - the unconditional coverage property places a restriction on how often VaR violations may occur. The independence property places a strong restriction on the ways in which these violations may occur. Specifically, any two elements of the hit sequence, $(I_{t+j}(\alpha), I_{t+k}(\alpha))$ must be independent from each other. Intuitively, this condition requires that the previous history of VaR violations, $\{\dots, I_{t-1}(\alpha), I_t(\alpha)\}$, must not convey any information about whether or not an additional VaR violation, $I_{t+1}(\alpha) = 1$, will occur. If previous VaR violations presage a future VaR violation then this points to a general inadequacy in the reported VaR measure. As an example, suppose that 1% VaR violations always occur in pairs so that one 1% VaR violation is immediately followed by another. Accordingly, the probability of observing a loss in excess of the 1% VaR after one has already been observed is 100% and not 1%. Hence, the reported VaR does not accurately reflect the loss that can be expected to be exceeded 1% of the time. In this example, the reported 1% VaR would have to be increased, i.e. a larger loss is to be expected following a VaR violation, in order to satisfy the independence property. In general, a clustering of VaR violations represents a violation of the independence property that signals a lack of responsiveness in the reported VaR measure as changing market risks fail to

be fully incorporated into the reported VaR measure thereby making successive runs of VaR violations more likely.

It is important to recognize that the unconditional coverage and independence properties of the hit sequence are separate and distinct and must both be satisfied by an accurate VaR model. In principle, a particular VaR model could result in a hit sequence that satisfies the unconditional coverage property but not the independence property. Likewise, another VaR model could result in a hit sequence that satisfies the independence property but not the unconditional coverage property. Only hit sequences that satisfy both properties can be described as evidence of an accurate VaR model. Each property characterizes a different dimension of an accurate VaR model.

In the case of the unconditional coverage property, it is clear that a VaR model which does not exhibit any predictability in violations but exhibits violations well in excess of the prescribed level, e.g. 1%, is problematic. Consider, for example, a bank that reports its actual VaR at the 10% level instead of the 1% level prescribed in the current regulatory framework. The independence property represents a more subtle yet equally important characteristic of an accurate VaR model. A VaR model that results in a predictable hit series implies that the underlying VaR model does not react quickly enough to changing market conditions. Consider a case in which VaR violations tend to cluster in groups. This clustering would imply that market risk capital requirements are underfunded for prolonged periods during episodes of increased risk. The consequences of being exposed to successive periods of heightened risk may be as problematic as a systematic under reporting of risk exposure. In particular, it may be even more difficult to recover from a string of unanticipated and large losses than it would be to recover from a higher than expected number of large losses that are spread out evenly over time. In this sense, it is important to recognize that the unconditional coverage and independence property jointly determine the accuracy of a given VaR model. A VaR model that satisfies one property or the other will result in an inaccurate description of the bank's risk exposure.

These two properties of the “hit” sequence, $[I_{t+1}(\alpha)]_{t=1}^{t=T}$, are often combined into the single statement,

$$I_t(\alpha) \stackrel{i.i.d.}{\sim} B(\alpha)$$

which reads that the hit sequence, $I_t(\alpha)$, is identically and independently distributed as a Bernoulli random variable with probability α . These two properties of the hit sequence

represent a complete characterization of an accurate *VaR* measure at a given level, α . Put differently, any series of *VaR* measures which are accurate in the sense that the reported *VaR* coincides with the α quantile of the profit and loss distribution must produce a hit series with the unconditional coverage and independence properties. Conversely, if a given *VaR* series produces a hit sequence that exhibits the unconditional coverage and independence properties then it is accurate in the sense previously described.

Ultimately, the main contribution of Christoffersen's (1998) insight is that the problem of determining whether a given *VaR* measure is accurate can be reduced to examining the unconditional coverage and independence properties of the resulting hit sequence. Accordingly, many of the *VaR* backtests that have been proposed in the last few years seek to examine whether the hit sequence satisfies one or both of these properties. We now turn to a discussion of these backtests. First, we discuss tests of the unconditional coverage property, then tests of the independence property and finally tests that seek to jointly examine the unconditional coverage and independence properties.

3 Tests of *VaR* Accuracy

3.1 Unconditional Coverage Tests

Some of the earliest proposed VaR backtests, e.g. Kupiec (1995), focused exclusively on the property of unconditional coverage. In short, these tests are concerned with whether or not the reported *VaR* is violated more (or less) than $\alpha \times 100\%$ of the time. Kupiec (1995), for example, proposed a proportion of failures or POF test that examines how many times a financial institution's *VaR* is violated over a given span of time. If the number of violations differs considerably from $\alpha \times 100\%$ of the sample, then the accuracy of the underlying risk model is called into question. Using a sample of T observations, Kupiec's (1995) test statistic takes the form,

$$\begin{aligned}
 POF &= 2 \log \left(\left(\frac{1 - \hat{\alpha}}{1 - \alpha} \right)^{T - I(\alpha)} \left(\frac{\hat{\alpha}}{\alpha} \right)^{I(\alpha)} \right) \\
 \hat{\alpha} &= \frac{1}{T} I(\alpha) \\
 I(\alpha) &= \sum_{t=1}^T I_t(\alpha)
 \end{aligned} \tag{5}$$

and close inspection of the test statistic reveals that if the proportion of VaR violations, $\hat{\alpha} \times 100\%$, is exactly equal to $\alpha \times 100\%$ then the POF test takes the value zero, indicating no evidence of any inadequacy in the underlying VaR measure. As the proportion of VaR violations differs from $\alpha \times 100\%$, the POF test statistic grows indicating mounting evidence that the proposed VaR measure either systematically understates or overstates the portfolio's underlying level of risk. In the case of a 1% VaR , if the test is carried out using one year, i.e. $T = 255$, of daily data then a common threshold for determining that a VaR measure is under reporting the actual 1% VaR is 8 violations or roughly 3% of the sample.²

At this point is useful to draw a connection between tests of unconditional coverage and the current regulatory framework. Recall that the market risk capital multiplier, S_t , is solely determined by the number of times the 1% VaR has been violated in the past 250 trading days. Similarly, Kupiec's POF test is only a function of the number of VaR violations during the sample period, $\hat{\alpha}$. Accordingly, there is a close connection between movements in the market risk multiplier and the POF test of unconditional coverage. For example, consider the fact that the market risk capital requirement begins to increase after four violations of the 1% VaR have occurred in the previous 250 days. This threshold is consistent with a value of the Kupiec POF test of 0.76. A realization of the POF test of this magnitude or larger would be expected to occur roughly 38% of the time in the case that the VaR model under consideration is in fact accurate. At the other extreme, observing ten VaR violations, in which case the multiplier is set to its maximal value of 4.0 and the underlying VaR model is deemed inaccurate requiring immediate steps to improve its accuracy, is equivalent to a POF test value of 12.95 which would be expected to occur much less than 1% of the time in the case that the VaR model under consideration is accurate. Accordingly, the market risk capital multiplier can be interpreted as a test of unconditional coverage that mandates a larger market risk capital set-aside as the evidence that the VaR model under consideration is inaccurate mounts.

Kupiec's POF test of unconditional coverage is a well known example of a VaR back-test. There are, however, a variety of statistical tests that could be employed to assess the unconditional coverage property of a given VaR model. One alternative would be to simply base a test directly on the sample average of the number of VaR violations over a given time period, $\hat{\alpha}$. Under the assumption that the VaR under consideration is accurate then a

²This is the threshold that is consistent with a Type I error rate of 5%. Put differently, if this threshold is used in judging the adequacy of a proposed VaR measure, the test will identify an accurate VaR measure as being inaccurate 5% of the time.

scaled version of $\hat{\alpha}$,

$$z = \frac{\sqrt{T}(\hat{\alpha} - \alpha)}{\sqrt{\alpha(1 - \alpha)}}, \quad (6)$$

has an approximate standard normal distribution. Moreover, the exact finite sample distribution of z is known and so hypothesis tests can be conducted in exactly the same way that hypothesis tests are conducted in the case of Kupiec's POF statistic.³

While tests of unconditional coverage provide a useful benchmark for assessing the accuracy of a given *VaR* model, these tests are hampered by two shortcomings. The first shortcoming is that these tests are known to have difficulty detecting *VaR* measures that systematically under report risk. From a statistical point of view these tests exhibit low power in sample sizes consistent with the current regulatory framework, i.e. one year. Kupiec (1995), for example, reports that when the threshold of 8 violations in one year is used in testing then the chance of detecting a *VaR* measure that reports the 3% *VaR* instead of the required 1% *VaR*, thereby under reporting its portfolio risk and under funding its market risk capital, is only 65%. In this case the odds of detecting an institution that systematically under reports their *VaR* is slightly better than a coin flip. Moreover, the size of the under reporting can be quite substantial. Consider a case in which portfolio profits and losses are normally distributed. In this situation, reporting the portfolio's 3% *VaR* instead of its 1% *VaR* would result in a capital charge that is 20% too small. Specifically, the 1% quantile of the normal distribution corresponds to 2.33 standard deviations while the 3% quantile corresponds to 1.88 standard deviations. If portfolio returns are more prone to very large profits and losses, as in the case of the *t* distribution, then the understatement will be even more severe. In the case that portfolio returns are distributed according to a *t*(5) distribution a similar calculation reveals that reporting the 3% *VaR* instead of the required 1% *VaR* would result in a capital charge that is 27% too small. The problem of low power may be ameliorated to some extent by increasing the sample over which the test is constructed. In the case of the above example, Kupiec (1995) reports that increasing the sample period from one year to two years increases the chance of detecting the systematic under reporting of *VaR* from 65% to roughly 90%. This increase in detection, however, comes at the cost of reducing the frequency with which the adequacy of the *VaR* measure can be assessed.

A second shortcoming of these tests is that they focus exclusively on the unconditional

³The z statistic is actually the Wald variant of the likelihood ratio statistic proposed by Kupiec (1995). One potential advantage of the Wald test over the likelihood ratio test is that it is well-defined in the case that no *VaR* violations occur. Kupiec's POF test is undefined in this case since the log of 0 is undefined. Moreover, the possibility that no violations occur in a period as short as one year is not trivial. Accordingly, it may be advisable to employ an unconditional coverage test which is well-defined in this event.

coverage property of an adequate *VaR* measure and do not examine the extent to which the independence property is satisfied. Accordingly, these tests may fail to detect *VaR* measures that exhibit correct unconditional coverage but exhibit dependent *VaR* violations. As discussed previously, *VaR* models that violate the independence property may result in losses that exceed the reported *VaR* in clusters or streaks. A streak of large unexpected losses may result in even more stress on a financial institution than large unexpected losses that occur somewhat more frequently than expected but are spread out over time. As an example, over a one year period, four losses in a row that exceed an institutions 1% *VaR* may be a larger signal of inadequate risk management practices than eight losses that occur evenly over the two year period. Moreover, it is interesting to note that while the second scenario would result an increase in the institutions market risk capital multiplier, the first scenario would not. To the extent that dependent *VaR* violations signal a lack of responsiveness to changing market conditions and inadequate risk reporting, relying solely on tests of unconditional coverage in backtesting would appear problematic.

3.2 Independence Tests

In light of the failure of tests of unconditional coverage to detect violations of the independence property of an accurate *VaR* measure, a variety of tests have been developed which explicitly examine the independence property of the *VaR* hit series, $I_t(\alpha)$. An early and influential test in this vein is Christoffersen's (1998) Markov test. The Markov test examines whether or not the likelihood of a *VaR* violation depends on whether or not a *VaR* violation occurred on the previous day. If the *VaR* measure accurately reflects the underlying portfolio risk then the chance of violating today's *VaR* should be independent of whether or not yesterday's *VaR* was violated. If, for example, the likelihood of a 1% *VaR* violation increased on days following a previous 1% *VaR* violation then this would indicate that the 1% *VaR* following a violation should in fact be increased. The test is carried out by creating a 2×2 contingency table that records violations of the institution's *VaR* on adjacent days as in Table 1. If the *VaR* measure accurately reflects the portfolio's risk then the proportion of violations that occur after a previous violation, $I_{t-1} = 1$, should be the same as the proportion of violations that occur after a day in which no violation occurred, $I_{t-1} = 0$. In terms of the elements in Table 1 it should be the case that $\frac{N_1}{N_1+N_3} = \frac{N_2}{N_2+N_4}$. If these proportions differ greatly from each other then this calls the validity of the *VaR* measure into question.

A more recent independence test that has been suggested by Christoffersen and Pelletier (2004) uses the insight that if *VaR* violations are completely independent from each other then the amount of time that elapses between *VaR* violations should be independent of the amount of time that has elapsed since the last violation. In this sense, the time between

VaR violations should not exhibit any kind of “duration dependence”. The chance that a *VaR* violation occurs in, say, the next 10 days should not depend upon whether the last *VaR* violation occurred 10 or 100 days ago. Unlike, the Markov test, this duration based test can not be constructed simply by computing a 2×2 contingency table. Carrying out the test requires estimating a statistical model for the duration of time between *VaR* violations by the method of maximum likelihood which must be done using numerical methods. Despite the lack of transparency in the test statistic, Christoffersen and Pelletier (2004) provide some evidence that this test of independence has more power than the Markov test to detect a *VaR* measure that violates the independence property. They show, for example, that in a setting in which the Markov test identifies an inaccurate *VaR* measure 28% of the time their duration based test identifies the inaccurate *VaR* measure 46% of the time.

While independence tests provide an important source of discriminatory power in detecting inadequate *VaR* measures, they are subject to one main drawback. All independence tests start from the assertion that any accurate *VaR* measure will result in a series of independent hits, $[I_t(\alpha)]_{t=1}^{t=T}$. Accordingly, any test of the independence property must fully describe the way in which violations of the independence property may arise. In the case of the Markov test, for example, the independence property may be violated by allowing for the possibility that the chance of violating tomorrow’s *VaR* depends on whether or not yesterday’s *VaR* was violated. There are, however, myriad ways in which the independence property might be violated. For example, it might be the case that the likelihood of violating tomorrow’s *VaR* depends not on whether yesterday’s *VaR* was violated but whether the *VaR* was violated one week ago. If this is the way in which the lack of the independence property manifests itself then the Markov test will have no ability to detect this violation of the independence property.

In statistical terms, any independence test must completely specify the alternative hypothesis that the independence property is being tested against. Intuitively, an independence test must describe the types of anomalies that it is going to look for when examining whether or not the independence property is satisfied. Violations of the independence property which are not related to the set of anomalies defined by the test will not be systematically detected by the test. As a result, independence tests are only likely to be effective at detecting inaccurate *VaR* measures to the extent that the tests are designed to identify violations of the independence property in ways that are likely to arise when internal risk models fail to provide accurate *VaR* measures. This kind of information would likely come from a thorough understanding of the types of situations in which common risk models fail to accurately describe portfolio risk. To the extent that risk models can be slow to react to changing market conditions, tests that examine the amount of clustering in *VaR* violations such as

the Markov test may be useful in identifying inaccurate *VaR* models.

3.3 Joint Tests of Unconditional Coverage and Independence

An accurate *VaR* measure must exhibit both the independence and unconditional coverage property. Accordingly, tests that jointly examine the unconditional coverage and independence properties provide an opportunity to detect *VaR* measures which are deficient in one way or the other. Both the Markov test of Christoffersen (1998) and the duration test of Christoffersen and Pelletier (2004) can be extended to jointly test for both independence and unconditional coverage. In the case of the Markov test, it is particularly simple to characterize how the joint test examines both properties of the *VaR* measure.

Recall that the Markov test proceeds by constructing a 2×2 contingency table that records the frequency of *VaR* violations and non-violations on successive days as in Table 1. The Markov independence test then examines whether the proportion of violations following previous violations is equal to the proportion of violations following previous non-violations. In terms of the notation in Table 1 this is equivalent to examining whether $\frac{N_1}{N_1+N_3} = \frac{N_2}{N_2+N_4}$. If the *VaR* measure also exhibits the unconditional coverage property then these proportions should match the total proportion of violations, $\frac{N_1+N_2}{N}$, and this should be identical to α , i.e. $\frac{N_1}{N_1+N_3} = \frac{N_2}{N_2+N_4} = \frac{N_1+N_2}{N} = \alpha$. Accordingly, the joint Markov test examines whether there is any difference in the likelihood of a *VaR* violation following a previous *VaR* violation or non-violation and simultaneously determines whether each of these proportions is significantly different from α .

The discussion of joint tests so far might seem to suggest that joint tests are universally preferable to tests of either the unconditional coverage property or independence property alone. While joint tests have the property that they will eventually detect a *VaR* measure which violates either of these properties, this comes at the expense of a decreased ability to detect a *VaR* measure which only violates one of the two properties. If, for example, a *VaR* measure exhibits appropriate unconditional coverage but violates the independence property then an independence test has a greater likelihood of detecting this inaccurate *VaR* measure than a joint test. The joint test is hampered by the fact that one of the two violations it is designed to detect, namely violations of unconditional coverage, is actually satisfied by the *VaR* measure. The fact that one of the two properties is satisfied makes it more difficult for the joint test to detect the inadequacy of the *VaR* measure. As an example, consider a 5% *VaR* measure which exhibits the unconditional coverage property but not the independence property. Specifically, suppose that a *VaR* violation occurs 20% of the time after a previous *VaR* violation but that a *VaR* violation only occurs 4.2% of the time after a period in which

no *VaR* violation occurs.⁴ In a one year sample of daily observations the joint Markov test detects this inaccurate *VaR* measure 50% of the time while the independence Markov test detects this inaccurate *VaR* measure 56% of the time. The increased power of detection that comes from choosing a test that focuses one of the two properties indicates that either unconditional coverage or independence tests alone are preferable to joint tests when prior considerations are informative about the source of the *VaR* measure's potential inaccuracy.

3.4 Tests Based On Multiple *VaR* Levels- α

All of the backtests discussed so far have focused on determining the adequacy of a *VaR* measure at a single level, α . In general, however, there is no need to restrict attention to a single *VaR* level. The unconditional coverage and independence property of an accurate *VaR* measure should hold for any level of α . Crnkovic and Drachman (1997), Diebold, Gunther and Tay (1998) as well as Berkowitz (2001) have all suggested backtests based on multiple *VaR* levels. These authors exploit the insight that if portfolio risk is adequately modeled then the 1% *VaR* should be violated 1% of the time, the 5% *VaR* should be violated 5% of the time, the 10% *VaR* should be violated 10% of the time and so on. Furthermore, a *VaR* violation at any level should be independent from a *VaR* violation at any other level so that, for example, a violation of a portfolio's 5% *VaR* today should not portend a violation of the portfolio's 1% *VaR* tomorrow. In short, *VaR* violations at all levels should be independent from each other.

Crnkovic and Drachman (1997) as well as Diebold, Gunther and Tay (1998) point out that this insight can be formalized in the following manner. Again, consider $x_{t,t+1}$ to be the realized profit or loss from time period t to $t + 1$, i.e. over one day. Next consider the quantile of the probability distribution that corresponds to the observation, $x_{t,t+1}$. In particular define,

$$z_{t+1} = F^{-1}(x_{t,t+1}|\Omega_t) \tag{7}$$

where, as before, $F^{-1}(\cdot|\Omega_t)$ refers to the conditional quantile function of the profit and loss distribution. It is worth comparing the observed quantile, z_t , to the hit indicator, $I_t(\alpha)$. The reported quantile provides a quantitative and continuous measure of the magnitude of realized profits and losses while the hit indicator only signals whether a particular threshold was exceeded. In this sense, the series of reported quantiles, $[z_t]_{t=1}^T$, provides more information about the accuracy of the underlying risk model. Moreover, a series of reported

⁴These relative frequencies imply that a *VaR* violation occurs 5% of the time on average.

quantiles that accurately reflects the actual distribution of profits and losses exhibits two key properties.

1. Uniformity - The series, $[z_t]_{t=1}^T$, should be uniformly distributed over the unit interval $[0, 1]$. This property of z_t is analogous to the statement that the $\alpha \times 100\%$ *VaR* should be violated $\alpha \times 100\%$ of the time for each and every α and is a direct parallel to the unconditional coverage property of the hit series, $I_t(\alpha)$, in the case of a single *VaR* measure.
2. Independence - The series, $[z_t]_{t=1}^T$, should be independently distributed. This is analogous to the statement that *VaR* violations should be independent from each other and that a *VaR* violation today at, say, the 1% level should not provide any information about whether or not a *VaR* violation will occur tomorrow at the 5% level. This property of the z_t series directly parallels the independence property that is required of the hit series, $I_t(\alpha)$, in the case of a single *VaR* measure.

These two properties are often combined into the single statement,

$$z_t \stackrel{i.i.d.}{\sim} U(0, 1) \tag{8}$$

which reads that the sequence of quantiles is identically and independently distributed as a Uniform random variable over the unit interval. A variety of statistical tests have been proposed to assess whether a reported series of quantiles, $[z_t]_{t=1}^T$, accord with the uniformity and independence properties. Just as in the case of assessing the accuracy of a single *VaR* measure these tests may be conducted individually, either the uniformity or independence property may be tested, or jointly.

The main advantage of these tests over and above tests based on a *VaR* measure at a single α level is that they, in principle, can provide additional power to detect an inaccurate risk model. By examining a variety of different quantiles, instead of a single quantile, these tests can detect violations of the independence or unconditional coverage property across a range of different *VaR* levels. This increased power to detect an inaccurate risk model, however, comes at some cost. One component of the cost comes in the form of an increased informational burden. In particular, in order to transform $x_{t,t+1}$ to z_{t+1} one must have access to the entire conditional quantile function, $F^{-1}(\cdot|\Omega_t)$, rather than a single quantile, $F^{-1}(\alpha|\Omega_t)$. Risk models that assume a particular form of the distribution of portfolio losses and gains may be reasonable models of extreme outcomes but may not be useful for characterizing the frequency of more moderate outcomes. If the underlying risk model is more focused on characterizing the stochastic behavior of extreme portfolio losses these models

may be misspecified over a range of profits and losses which are not highly relevant from a risk management perspective. Accordingly, tests that employ the entire series of quantiles, z_t , may signal an inaccurate model due to this source of misspecification.

A second component of the cost relates to a broadening of the definition of portfolio risk. Statistical tests that examine a variety of different quantiles implicitly broaden the definition of portfolio risk relative to tests that focus on a VaR measure at a single α level. Current regulatory guidelines dictate that portfolio risk be measured in terms of the 1% VaR for the purpose of determining risk-based capital requirements. Accordingly, examining the behavior of VaR measures at levels aside from the 1% level broadens the scope of the definition of portfolio risk. The extent to which the definition of portfolio risk is broadened, however, may be directly controlled by choosing to examine the uniformity and independence properties of the z series over a particular range of α levels.

One could, for example, rely on a test that examines the frequency of VaR violations over a range of predetermined α levels. An example of one such test is Pearson's Q test for goodness of fit. Pearson's Q test is based upon the number of observed violations at a variety of different VaR levels. The test is constructed as follows.⁵ First, partition the unit interval into different sub-intervals. For example, one could choose the partition $[0.00, 0.01]$, $[0.01, 0.05]$, $[0.05, 0.10]$, $[0.10, 1.00]$ which results in four separate regions or bins on the unit interval. Once a partition is chosen, simply count the number of VaR violations that occur within each bin. The number of VaR violations that occur in the $[0.00, 0.01]$ range, for example, records the number of days on which a loss in excess of the 1% VaR occurred. Likewise, the number of VaR violations that occur in the $[0.10, 1.00]$ range records the number of days on which a loss less extreme than the 10% VaR occurred. With the number of violations that have occurred within each bin in hand the Q test is computed according to the formula,

$$Q = \sum_{i=1}^k \frac{(N_{(l_i, u_i)} - N(u_i - l_i))^2}{N(u_i - l_i)} \quad (9)$$

where $N_{(l_i, u_i)}$ refers to the number of VaR violations in the i^{th} bin and N refers to the total number of days being used to construct the test. Also, l_i and u_i refer to the lower and upper bound of each bin. In the case that the VaR model being tested is accurate in the sense that the model's VaR coincides with actual VaR the test is approximately distributed according to the chi squared distribution with $k - 1$ degrees of freedom.

As the equation for Q makes clear, the particular partition of the unit interval that is

⁵Details concerning the foundations and derivation of this test statistic can be found in, DeGroot M. (1989), Probability and Statistics, Addison-Wesley, Reading, Massachusetts.

chosen may be tailored to focus on a particular set of quantiles of interest. If only quantiles more extreme than the 10th percentile are of interest, for example, one could use the partition [0.00, 0.01], [0.01, 0.05], [0.05, 0.10], [0.10, 1.00]. If only quantiles more extreme than the 5th percentile are of interest than the partition [0.00, 0.01], [0.01, 0.05], [0.05, 1.00] could be used. In this way the particular partition employed can be used as a means of controlling how broadly risk is defined in terms of the test.

In order to have a grasp of the quantitative nature of the tradeoff between increasing power and broadening the scope of the definition of risk, it is useful to examine an example. In order to better gauge this tradeoff an experiment is conducted in which profits and losses are generated over a one-year period from a statistical model. The profits and losses are then analyzed using three different risk models which are employed in the practice of risk management to varying degrees. Each of the three risk models examined is inaccurate in the sense that none of them accurately models the true underlying distribution of profits and losses. Each of the three models is then used to produce a one-year sample of observed quantiles, $[z_t]_{t=1}^{255}$. The quantiles are then used in conducting Pearson's Q test. The power of Pearson's Q test is compared to the power of Kupiec's (1995) test of unconditional coverage at the 1% level to assess the increase in power that arises from considering other quantiles besides the first percentile. Also, in addition to these three risk models we also consider the power of Pearson's Q test and Kupiec's (1995) test in the case that VaR is systematically under reported. We examine levels of under reporting that range between 5% - 25%.

The statistical model for portfolio profits and losses is one that assumes a fixed mean of the profit and loss distribution and focuses on a model of time varying volatility or risk. The model of time varying risk is one that incorporates three distinct features of volatility that have been recognized to be salient features of a variety of different financial asset markets ranging from stock markets to currency markets. First, the model provides for considerable variation in volatility. Each period volatility changes and sometimes changes quite rapidly so that a 25% change in volatility over a one month horizon is not particularly unlikely. Second, movements in volatility are assumed to exhibit a considerable degree of persistence. Much research on financial asset volatility finds that increases and decreases in volatility can be long lived. Volatility does tend towards a long run mean but shocks that either increase or decrease volatility often persist for several months. Lastly, the model implies that volatility tends to increase more after large portfolio losses than after large gains. The particular model for volatility employed in the experiment is Nelson's (1991) EGARCH(1,1) model. This model is widely used in empirical studies of financial market volatility and also bears a close resemblance to the model employed by Berkowitz and O'Brien (2002) in their empirical

study of bank trading revenues and VaR. The model's particular specification takes the form,

$$\begin{aligned}
x_{t,t+1} &= v_{t,t+1} \\
v_{t,t+1} &\sim N(0, \sigma_t^2) \\
\ln(\sigma_t^2) &= 0.02 + 0.94 \ln(\sigma_{t-1}^2) + 0.22 \left| \frac{v_{t-1,t}}{\sigma_{t-1}} \right| - 0.05 \left(\frac{v_{t-1,t}}{\sigma_{t-1}} \right)
\end{aligned} \tag{10}$$

where the specific model parameters have been chosen to be consistent with the behavior of monthly U.S. stock returns over the period 1927-1998. In particular note that the final term in the volatility specification, $-0.05 \left(\frac{v_{t-1,t}}{\sigma_{t-1}} \right)$, implies that volatility, $\ln(\sigma_t^2)$, rises more after a portfolio loss than a gain.

Three different models are used to measure portfolio risk ranging from naive to more sophisticated. The first and most naive model assumes, correctly, that losses and gains are normally distributed but assumes that the variance is constant. Accordingly, the *VaR* model uses the standard sample estimate of variance, $\hat{\sigma}_T^2 = \frac{\sum_{j=1}^{T-1} v_{j,j+1}^2}{T}$. With this variance estimate in hand, the profit or loss, $v_{t-1,t}$, on any given day can be sorted into its respective bin by computing,

$$[\sigma_{t-1} \Phi^{-1}(l_i), \sigma_{t-1} \Phi^{-1}(u_i)], \tag{11}$$

for each bin and then assigning the profit or loss to the bin that brackets $v_{t-1,t}$. As usual $\Phi^{-1}(\cdot)$ refers to the inverse of the standard normal distribution.

The second *VaR* model recognizes that volatility changes over time but is unaware of the particular way in which it changes over time. As a result, the model uses an exponentially weighted moving average of past squared profits and losses to estimate current volatility. The weighted moving average volatility estimate takes the form,

$$\sigma_{WMA,t}^2 = 0.97 \sigma_{RM,t}^2 + 0.03 v_{t-1,t}^2 \tag{12}$$

where the decay factor, 0.97, implies that volatility moves in a rather persistent fashion over time. The exact value of the decay factor was chosen to be consistent with the values that are employed in practice which often range between 0.94 and 0.97 (Simons, 1996). Again, the profit or loss on any day can be sorted into its respective bin by computing,

$$[\sigma_{WMA,t-1} \Phi^{-1}(l_i), \sigma_{WMA,t-1} \Phi^{-1}(u_i)],$$

for each bin and then assigning the profit or loss to the bin which brackets $v_{t-1,t}$.

The third model employs the Historical Simulation approach. Unlike the previous two models, the Historical Simulation approach makes no assumptions about the distribution of the profits and losses and does not explicitly model volatility. Instead, Historical Simulation uses the previous observed history of profits and losses and equates the distribution of future profits and losses with the observed empirical distribution, (\widehat{F}) .⁶ As is common in the implementation of the Historical Simulation method one year of data is first used in order to construct the empirical distribution function. Each realized profit or loss is associated with a bin by ranking it among the previous 255 days' trading experience. If, for example, today's loss was the 10th worst loss over the past 255 days then the loss would be associated with the bin that brackets the 4th percentile ($\frac{10}{255} = 0.039 \approx 0.04$). More details on the historical simulation method, its application and problems it presents for risk management can be found in Pritsker (2001).

The final scenario we consider to assess the value of Pearson's Q test as a backtesting procedure is a setting in which the true VaR is systematically under reported. Specifically, at each point in time the actual volatility of the future P&L distribution, σ_t , is known. The reported VaR, however, is based on only a fraction, β , of true volatility. Accordingly, any particular profit or loss, $v_{t-1,t}$, is sorted into one of the bins,

$$[(1 - \beta)\sigma_t\Phi^{-1}(l_i), (1 - \beta)\sigma_t\Phi^{-1}(u_i)],$$

and varying amounts of under reporting are considered ranging from no under reporting, $\beta = 0$, to a level of 25%, $\beta = 0.25$.

It is important to consider how this setting differs from the other three risk models. In each of the previous three models, VaR is sometimes overstated and sometimes understated depending on whether the estimated P&L volatility is above or below actual volatility. The presence of systematic under reporting, however, implies that the reported VaR is always below the actual VaR. Accordingly, this scenario may more realistically capture an environment in which a bank's risk model consistently fails to account for a portion of the risk inherent in its trading activity.

Tables 2 and 3 display the power of Pearson's Q test and Kupiec's (1995) test of unconditional coverage at the 1% level to identify each of the three risk models as well as the systematically under reported VaR as inaccurate. The power is computed by simulating the statistical model for the P&L for one year and then computing both Pearson's Q and Kupiec's (1995) backtest. This process was repeated 1,000 times and the fraction of test statistics that exceed the 95% critical value is the estimate of the test's power. In the case

⁶The empirical CDF, \widehat{F} in a sample of T observations is defined as $\widehat{F}(c) = \frac{1}{T} \sum_{t=1}^T 1(x_t \leq c)$.

of Pearson's Q test, the partition, $[0.00, 0.01]$, $[0.01, 0.05]$, $[0.05, 0.10]$, $[0.10, 1.00]$ was chosen to focus on percentiles that are associated with relatively rare occurrences but does expand the scope of the definition of risk beyond the first percentile.

Table 2 reports the power of the Q test to identify each of the first three risk models, recursive, weighted moving average and historical simulation, as inaccurate. The table suggests that with respect to the first three risk models, some modest gains in power are attainable by expanding the range of quantiles that are analyzed in the backtest. In the case of the most inaccurate VaR model, the recursive model, Pearson's Q test identified it as inaccurate 33.3 percent of the time while the Kupiec (1995) backtest only identifies the recursive model in 25.5 percent of all simulations. The power gains are less impressive for VaR models that are less egregiously misspecified. In the case of historical simulation, the Q test only marginally increases power from 11.8% to 14.8%. The gain in the case of the moving weighted average model is imperceptible as both tests identify the model as inaccurate in 5% of the simulations. The inability of either tests to identify the weighted moving average model as inaccurate is due to the fact that while this model does not perfectly model P&L risk, it provides a very close approximation to the underlying level of risk. These results suggest that the power gains may be largest in cases where the degree of misspecification is largest which is noteworthy since these are precisely the cases in which identifying an inadequate risk model may be most important.

Table 3 displays the power of Pearson's Q test and the Kupiec (1995) backtest in the case that VaR is systematically under reported. The table displays the power results for varying amounts of under reporting ranging between 5%-25%. The last row of the table reports the ratio of the power of the Pearson Q to the Kupiec backtest. At lower levels of systematic under reporting, between 5% and 10%, the benefit of Pearson's Q test are largest in terms of proportional increase in power. At the 5% level of under reporting, for instance, Pearson's Q test identifies the VaR model as being inaccurate 13.5% of the time as opposed to only 6.3% of the time in the case of the Kupiec backtest. The proportional gain in power declines as the degree of under reporting increases and the power of both tests rises towards unity. In the case of a 25% under reporting in VaR, for example, the Q test only provides a 10% improvement in power over and above the Kupiec test. Despite the smaller proportional increase, however, the absolute gain in power is substantial. In the case of a 25% under reporting of VaR, Pearson's Q identifies the model as inaccurate 94.2% of the time as opposed to 79.7% of the time in the case of the Kupiec test. These results suggest that Pearson's Q test may provide more substantial power gains in settings where there risk is systematically under reported. While the Q test does exhibit some ability to increase power in the case of the other three risk models, the gain in power in the case of systematic

under reporting is considerably larger.

4 Loss Function Based Backtests

All of the backtests that have been discussed so far have focused on examining the behavior of the hit function. Specifically, recall that the hit function is defined as follows,

$$I_{t+1}(\alpha) = \begin{cases} 1 & \text{if } x_{t,t+1} \leq -VaR_t(\alpha) \\ 0 & \text{if } x_{t,t+1} > -VaR_t(\alpha) \end{cases}$$

where $x_{t,t+1}$ represents the profit or loss between the end of day t and $t + 1$. Despite the fact that the hit function plays a prominent role in a variety of backtesting procedures, the information contained in the hit function is limited. One might be interested, for example, in the magnitude of the exceedance rather than simply whether or not an exceedance occurred. In general, a backtest could be based on a function of the observed profit or loss and the corresponding model VaR . This would result in the construction of a general loss function, $L(VaR_t(\alpha), x_{t,t+1})$, which could be evaluated using past data on profits and losses and the reported VaR series. Lopez (1999a,b) suggests this approach to backtesting as an alternative to the approach that focuses exclusively on the hit series.

As an example, consider a loss function which measures the difference between the observed loss and the VaR in cases where the loss exceeds the reported VaR measure. Specifically, consider the loss function suggested by Lopez (1999b),

$$L(VaR_t(\alpha), x_{t,t+1}) = \begin{cases} 1 + (x_{t,t+1} - VaR_t(\alpha))^2 & \text{if } x_{t,t+1} \leq -VaR_t(\alpha) \\ 0 & \text{if } x_{t,t+1} > -VaR_t(\alpha) \end{cases} \quad (13)$$

which measures how well the model VaR predicts losses when they occur. A backtest could then be conducted by using actual data on reported VaR and profits and losses to assess whether or not the realized losses are in accord with what would be expected when the reported VaR accurately reflects underlying risk. In particular, a backtest using this loss function based approach would typically be based on the sample average loss,

$$\widehat{L} = \frac{1}{T} \sum_{t=1}^T L(VaR_t(\alpha), x_{t,t+1}), \quad (14)$$

and then the degree to which the observed average loss is consistent with an accurate VaR model would have to be assessed. This example highlights the increased flexibility of the

loss function based approach. The loss function may be tailored to address specific concerns that may be of interest when analyzing the performance of a *VaR* model.

Unfortunately this increased flexibility comes at a substantial increase in the informational burden associated with assessing the accuracy of the *VaR* model under consideration. Specifically, in order to determine whether the average loss, \widehat{L} , is “too large relative to what would be expected” it is necessary to understand the stochastic behavior of the loss function, $L(VaR(\alpha), x_{t,t+1})$. Generally, this would require that the backtest make an explicit assumption about the distribution of profits and losses, $x_{t,t+1}$. Lopez (1999b) suggests a three step procedure for determining the range of values for average loss that are consistent with an accurate *VaR* model.

1. Fit a statistical model to the underlying profit and loss data and determine a suitable model of the distribution of profits and losses, $f(x_{t,t+1}|\Omega_t)$.
2. Generate, from the model derived in the first step, a history of profits and losses and the associated *VaR*, $VaR_t(\alpha)$, and construct a value for the average loss, \widehat{L}_i .
3. Repeat the above process for a very large number of trials, say 10,000, and use the resulting set of average losses, $\left[\widehat{L}_i\right]_{i=1}^{i=10,000}$ as an estimate of the distribution of the average loss.

The quantiles of the empirical distribution of the simulated average losses, $\left[\widehat{L}_i\right]_{i=1}^{i=10,000}$, may be used just as in a standard hypothesis testing framework to determine when a proposed *VaR* model is acceptably accurate.

As the description of this method makes clear, the backtesting procedure relies on a precise description of the stochastic behavior of profits and losses, $f(x_{t,t+1}|\Omega_t)$. As a result, a finding that the average observed loss is “too large” could either signal an inaccurate risk model or an inaccurate assumption about the stochastic behavior of profits and losses. This feature of the loss function based approach is at odds with tests that examine the stochastic properties of the hit function, $I_{t+1}(\alpha)$. Recall that regardless of the stochastic properties of profits and losses, $f(x_{t,t+1}|\Omega_t)$, as long as the underlying *VaR* model accurately reflects the true risk exposure, the hit function has the property that it is independently and identically distributed as a Bernoulli random variable with a probability of success of precisely $\alpha \times 100\%$. Accordingly, inferences about the accuracy of the *VaR* model may be made without assuming anything about the stochastic behavior of profits and losses. This is a property of hit function based backtests which is not shared by the more general loss function based backtests. This is not to say, however, that loss function based backtests have no place in the backtesting process. Loss function based backtests, for example, may be

extremely useful for determining whether one *VaR* model provides a better risk assessment than another competing *VaR* model. In this situation, the loss function could be tailored to address specific concerns about how each *VaR* model behaves in different types of market conditions. In this sense, loss function based backtests may be more suited to discriminating among competing *VaR* models rather than judging the accuracy of a single model.

5 Conclusion

Verifying the accuracy of internal risk models used in setting market risk capital requirements requires backtesting. The current regulatory framework uses a “traffic light” approach to backtesting that is related to Kupiec’s (1995) proportion of failures test. In this paper, the fundamental properties of an accurate VaR model, independence and unconditional coverage were defined and their relevance from a risk management perspective was discussed. A variety of extant tests that examine the validity of the independence property, the unconditional coverage property or both properties were reviewed. Backtests that use pre-specified loss functions to evaluate the accuracy of a VaR model were also reviewed. Also, a test that focuses on other quantiles of the P&L distribution besides the first percentile was discussed and its power properties were examined in a small simulation experiment. The results of the simulation experiment suggest that moderate gains in statistical power relative to the power of a test that only examines the first percentile can be achieved by examining other quantiles besides the first percentile. Tests that examine several quantiles are most successful in identifying inaccurate VaR models in the presence of systematic under reporting of risk.

BIBLIOGRAPHY

- Berkowitz, J., "Testing Density Forecasts With Applications to Risk Management," *Journal of Business and Economic Statistics*, 19, 2001, 465-474.
- Berkowitz, J. and O'Brien J., "How Accurate Are the Value-at-Risk Models at Commercial Banks," *Journal of Finance*, 57, 2002, 1093-1112.
- Christoffersen P., "Evaluating Interval Forecasts," *International Economic Review*, 39, 1998, 841-862.
- Christoffersen P., and Pelletier D., "Backtesting Value-at-Risk: A Duration-Based Approach," *Journal of Empirical Finance*, 2, 2004, 84-108.
- Cassidy, C. and Gizycki M., "Measuring Traded Market Risk: Value-At-Risk And Backtesting Techniques," Research Discussion Paper 9708, Reserve Bank of Australia, 1997.
- Crnkovic C., and Drachman J., "Quality Control," in VaR: Understanding and Applying Value-at-Risk, 1997, London, Risk Publications.
- DeGroot M. (1989), Probability and Statistics, Addison-Wesley, Reading, Massachusetts.
- Diebold F.X., Gunther T., and Tay A., "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, 39, 1998, 863-883.
- Haas M., "New Methods in Backtesting," Financial Engineering Research Center, Working Paper, 2001.
- Kerkhoff J., and Melenberg B., "Backtesting for Risk-Based Regulatory Capital," Tilburg University, Working Paper, 2003.
- Kupiec P., "Techniques for Verifying the Accuracy of Risk Management Models," *Journal of Derivatives*, 3, 1995, 73-84.
- Lopez J.A., "Regulatory Evaluation of Value-at-Risk Models," *Journal of Risk*, 1999a, 1, 37-64.
- Lopez J.A., "Methods for Evaluating Value-at-Risk Models," *Federal Reserve Bank of San Francisco Economic Review*, 1999b, 2, 3-17.
- Pritsker M., "Evaluating Value-at-Risk Methodologies: Accuracy versus Computational Time," *Journal of Financial Services Research*, 1997, 12, 201-241.
- Pritsker M., "The Hidden Dangers of Historical Simulation," Working Paper 2001-27, Board of Governors of the Federal Reserve System, 2001.
- Simons K. (1996), "Value at Risk - New Approaches to Risk Management," VAR: Understanding and Applying Value-at-Risk, RISK publications, London.
- Stahl G., "Backtesting Using a Generalisation of the Traffic-Light-Approach," *Neural Network World*, 1997, 4, 565-577.
- Sullivan, J., Brooks, R., and Stoumbos Z., "Assessing the Accuracy of Value at Risk," Working Paper Rutgers University, 2003.

Table 1: Contingency Table for Markov Independence Test

	$I_{t-1} = 0$	$I_{t-1} = 1$	
$I_t = 0$	N_1	N_2	$N_1 + N_2$
$I_t = 1$	N_3	N_4	$N_3 + N_4$
	$N_1 + N_3$	$N_2 + N_4$	N

Table 2: Power of Pearson's Q and Kupiec Tests (%)

<u>Risk Model</u>	<u>Q Test</u>	<u>Kupiec Test</u>
Recursive	33.3	25.5
Weighted Moving Average	5.00	4.90
Historical Simulation	14.8	11.8

Table 3: Power of Pearson's Q and Kupiec Tests (%)

<u>Under Reporting Level (α)</u>	<u>5%</u>	<u>10%</u>	<u>15%</u>	<u>20%</u>	<u>25%</u>
<u>backtest</u>					
Pearson's Q	13.5	35.9	63.8	86.0	94.2
Kupiec Test	6.30	19.4	43.8	69.0	79.7