CHAPTER 6

Backtesting VAR

> Disclosure of quantitative measures of market risk, such as value-at-risk, is enlightening only when accompanied by a thorough discussion of how the risk measures were calculated and how they related to actual performance.
>
> —*Alan Greenspan (1996)*

Value-at-risk (VAR) models are only useful insofar as they predict risk reasonably well. This is why the application of these models always should be accompanied by validation. *Model validation* is the general process of checking whether a model is adequate. This can be done with a set of tools, including backtesting, stress testing, and independent review and oversight.

This chapter turns to backtesting techniques for verifying the accuracy of VAR models. *Backtesting* is a formal statistical framework that consists of verifying that actual losses are in line with projected losses. This involves systematically comparing the history of VAR forecasts with their associated portfolio returns.

These procedures, sometimes called *reality checks*, are essential for VAR users and risk managers, who need to check that their VAR forecasts are well calibrated. If not, the models should be reexamined for faulty assumptions, wrong parameters, or inaccurate modeling. This process also provides ideas for improvement and as a result should be an integral part of all VAR systems.

Backtesting is also central to the Basel Committee's ground-breaking decision to allow internal VAR models for capital requirements. It is unlikely the Basel Committee would have done so without the discipline of a rigorous backtesting mechanism. Otherwise, banks may have an incentive

to understate their risk. This is why the backtesting framework should be designed to maximize the probability of catching banks that willfully understate their risk. On the other hand, the system also should avoid unduly penalizing banks whose VAR is exceeded simply because of bad luck. This delicate choice is at the heart of statistical decision procedures for backtesting.

Section 6.1 provides an actual example of model verification and discusses important data issues for the setup of VAR backtesting. Next, Section 6.2 presents the main method for backtesting, which consists of counting deviations from the VAR model. It also describes the supervisory framework by the Basel Committee for backtesting the internal-models approach. Section 6.3 illustrates practical uses of VAR backtesting.

6.1 SETUP FOR BACKTESTING

VAR models are only useful insofar as they can be demonstrated to be reasonably accurate. To do this, users must check systematically the validity of the underlying valuation and risk models through comparison of predicted and actual loss levels.

When the model is perfectly calibrated, the number of observations falling outside VAR should be in line with the confidence level. The number of exceedences is also known as the number of *exceptions*. With too many exceptions, the model underestimates risk. This is a major problem because too little capital may be allocated to risk-taking units; penalties also may be imposed by the regulator. Too few exceptions are also a problem because they lead to excess, or inefficient, allocation of capital across units.

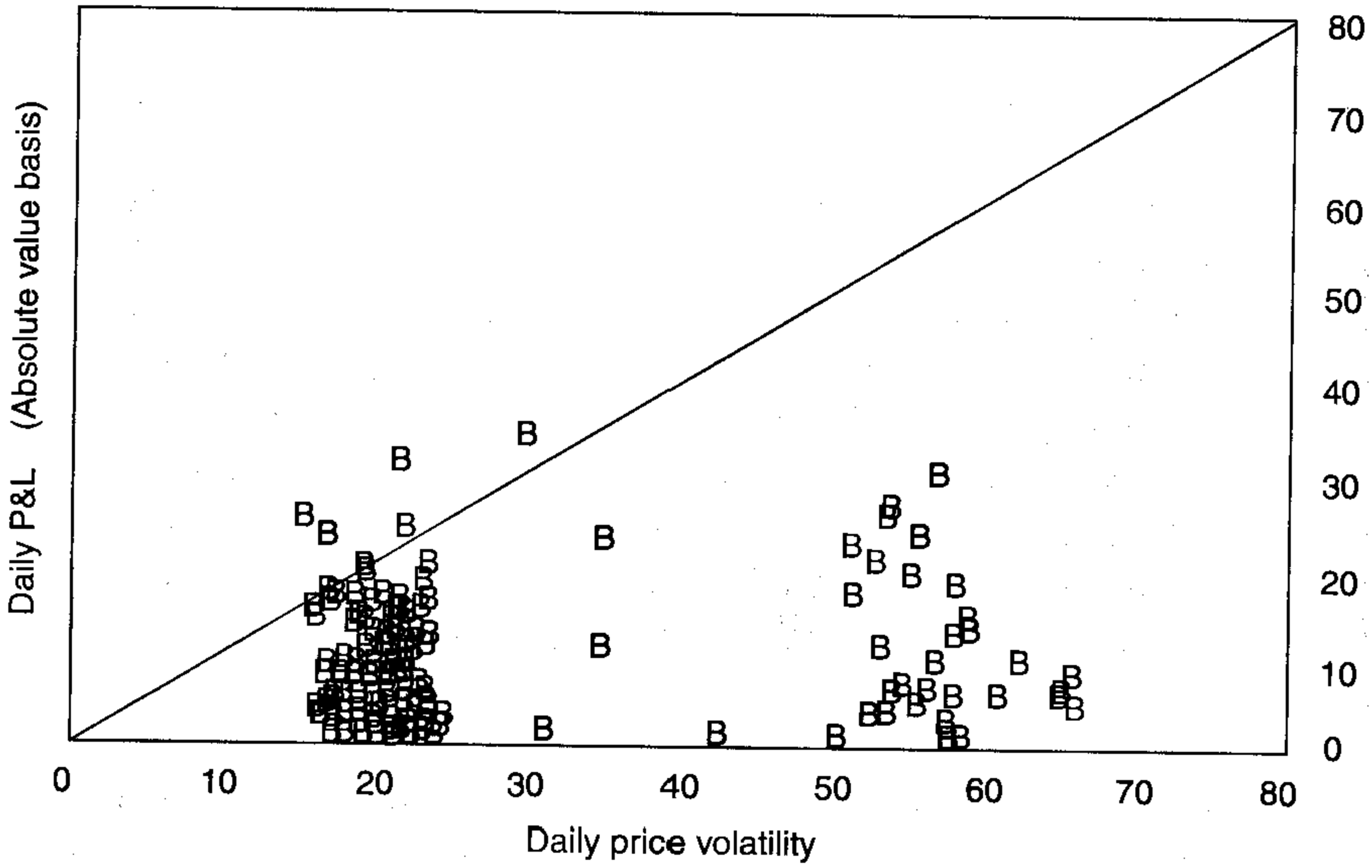
6.1.1. An Example

An example of model calibration is described in Figure 6-1, which displays the fit between actual and forecast daily VAR numbers for Bankers Trust. The diagram shows the absolute value of the daily profit and loss (P&L) against the 99 percent VAR, defined here as the *daily price volatility*.[1] The graph shows substantial time variation in the VAR measures, which

[1] Note that the graph does not differentiate losses from gains. This is typically the case because companies usually are reluctant to divulge the extent of their trading losses. This illustrates one of the benefits of VAR relative to other methods, namely, that by taking the absolute value, it hides the direction of the positions.

FIGURE 6-1

Model evaluation: Bankers Trust.



reflects changes in the risk profile of the bank. Observations that lie above the diagonal line indicate days when the absolute value of the P&L exceeded the VAR.

Assuming symmetry in the P&L distribution, about 2 percent of the daily observations (both positive and negative) should lie above the diagonal, or about 5 data points in a year. Here we observe four exceptions. Thus the model seems to be well calibrated. We could have observed, however, a greater number of deviations simply owing to bad luck. The question is: At what point do we reject the model?

6.1.2. Which Return?

Before we even start addressing the statistical issue, a serious data problem needs to be recognized. VAR measures assume that the current portfolio is "frozen" over the horizon. In practice, the trading portfolio evolves dynamically during the day. Thus the actual portfolio is "contaminated" by changes in its composition. The *actual return* corresponds to the actual P&L, taking into account intraday trades and other profit items such as fees, commissions, spreads, and net interest income.

This contamination will be minimized if the horizon is relatively short, which explains why backtesting usually is conducted on daily returns. Even so, intraday trading generally will increase the volatility of revenues because positions tend to be cut down toward the end of the trading day. Counterbalancing this is the effect of fee income, which generates steady profits that may not enter the VAR measure.

For verification to be meaningful, the risk manager should track both the actual portfolio return R_t and the hypothetical return R_t^* that most closely matches the VAR forecast. The *hypothetical return* R_t^* represents a frozen portfolio, obtained from fixed positions applied to the actual returns on all securities, measured from close to close.

Sometimes an approximation is obtained by using a *cleaned return*, which is the actual return minus all non-mark-to-market items, such as fees, commissions, and net interest income. Under the latest update to the *market-risk amendment*, supervisors will have the choice to use either hypothetical or cleaned returns.[2]

Since the VAR forecast really pertains to R^*, backtesting ideally should be done with these hypothetical returns. Actual returns do matter, though, because they entail real profits and losses and are scrutinized by bank regulators. They also reflect the true ex post volatility of trading returns, which is also informative. Ideally, both actual and hypothetical returns should be used for backtesting because both sets of numbers yield informative comparisons. If, for instance, the model passes backtesting with hypothetical but not actual returns, then the problem lies with intraday trading. In contrast, if the model does not pass backtesting with hypothetical returns, then the modeling methodology should be reexamined.

6.2 MODEL BACKTESTING WITH EXCEPTIONS

Model backtesting involves systematically comparing historical VAR measures with the subsequent returns. The problem is that since VAR is reported only at a specified confidence level, we expect the figure to be exceeded in some instances, for example, in 5 percent of the observations at the 95 percent confidence level. But surely we will not observe exactly

[2] See BCBS (2005b).

5 percent exceptions. A greater percentage could occur because of bad luck, perhaps 8 percent. At some point, if the frequency of deviations becomes too large, say, 20 percent, the user must conclude that the problem lies with the model, not bad luck, and undertake corrective action. The issue is how to make this decision. This *accept or reject decision* is a classic statistical decision problem.

At the outset, it should be noted that this decision must be made at some confidence level. The choice of this level for the *test*, however, is not related to the quantitative level p selected for VAR. The decision rule may involve, for instance, a 95 percent confidence level for backtesting VAR numbers, which are themselves constructed at some confidence level, say, 99 percent for the Basel rules.

6.2.1. Model Verification Based on Failure Rates

The simplest method to verify the accuracy of the model is to record the *failure rate*, which gives the proportion of times VAR is exceeded in a given sample. Suppose a bank provides a VAR figure at the 1 percent left-tail level ($p = 1 - c$) for a total of T days. The user then counts how many times the actual loss exceeds the previous day's VAR. Define N as the number of exceptions and N/T as the failure rate. Ideally, the failure rate should give an *unbiased* measure of p, that is, should converge to p as the sample size increases.

We want to know, at a given confidence level, whether N is too small or too large under the null hypothesis that $p = 0.01$ in a sample of size T. Note that this test makes no assumption about the return distribution. The distribution could be normal, or skewed, or with heavy tails, or time-varying. We simply count the number of exceptions. As a result, this approach is fully *nonparametric*.

The setup for this test is the classic testing framework for a sequence of success and failures, also called *Bernoulli trials*. Under the null hypothesis that the model is correctly calibrated, the number of exceptions x follows a *binomial* probability distribution:

$$f(x)=\binom{T}{x} p^x (1-p)^{T-x} \tag{6.1}$$

We also know that x has expected value of $E(x) = pT$ and variance $V(x) = p(1-p)T$. When T is large, we can use the central limit theorem and approximate the binomial distribution by the normal distribution

$$z = \frac{x - pT}{\sqrt{p(1-p)T}} \sim N(0,1) \tag{6.2}$$

which provides a convenient shortcut. If the decision rule is defined at the two-tailed 95 percent test confidence level, then the cutoff value of $|z|$ is 1.96. Box 6-1 illustrates how this can be used in practice.

This binomial distribution can be used to test whether the number of exceptions is acceptably small. Figure 6-2 describes the distribution when the model is calibrated correctly, that is, when $p = 0.01$ and with 1 year of data, $T = 250$. The graph shows that under the null, we would observe more than four exceptions 10.8 percent of the time. The 10.8 percent number describes the probability of committing a *type 1* error, that is, rejecting a correct model.

Next, Figure 6-3 describes the distribution of number of exceptions when the model is calibrated incorrectly, that is, when $p = 0.03$ instead of 0.01. The graph shows that we will not reject the incorrect model more than 12.8 percent of the time. This describes the probability of committing a *type 2* error, that is, not rejecting an incorrect model.

BOX 6-1

J.P. MORGAN'S EXCEPTIONS

In its 1998 annual report, the U.S. commercial bank J.P. Morgan (JPM) explained that

> In 1998, daily revenue fell short of the downside (95 percent VAR) band . . . on 20 days, or more than 5 percent of the time. Nine of these 20 occurrences fell within the August to October period.

We can test whether this was bad luck or a faulty model, assuming 252 days in the year. Based on Equation (6.2), we have $z = (x - pT)/\sqrt{p(1-p)T} = (20 - 0.05 \times 252)/\sqrt{0.05(0.95)252} = 2.14$. This is larger than the cutoff value of 1.96. Therefore, we reject the hypothesis that the VAR model is unbiased. It is unlikely (at the 95 percent test confidence level) that this was bad luck.

The bank suffered too many exceptions, which must have led to a search for a better model. The flaw probably was due to the assumption of a normal distribution, which does not model tail risk adequately. Indeed, during the fourth quarter of 1998, the bank reported having switched to a "historical simulation" model that better accounts for fat tails. This episode illustrates how backtesting can lead to improved models.

FIGURE 6-2

Distribution of exceptions when model is correct.

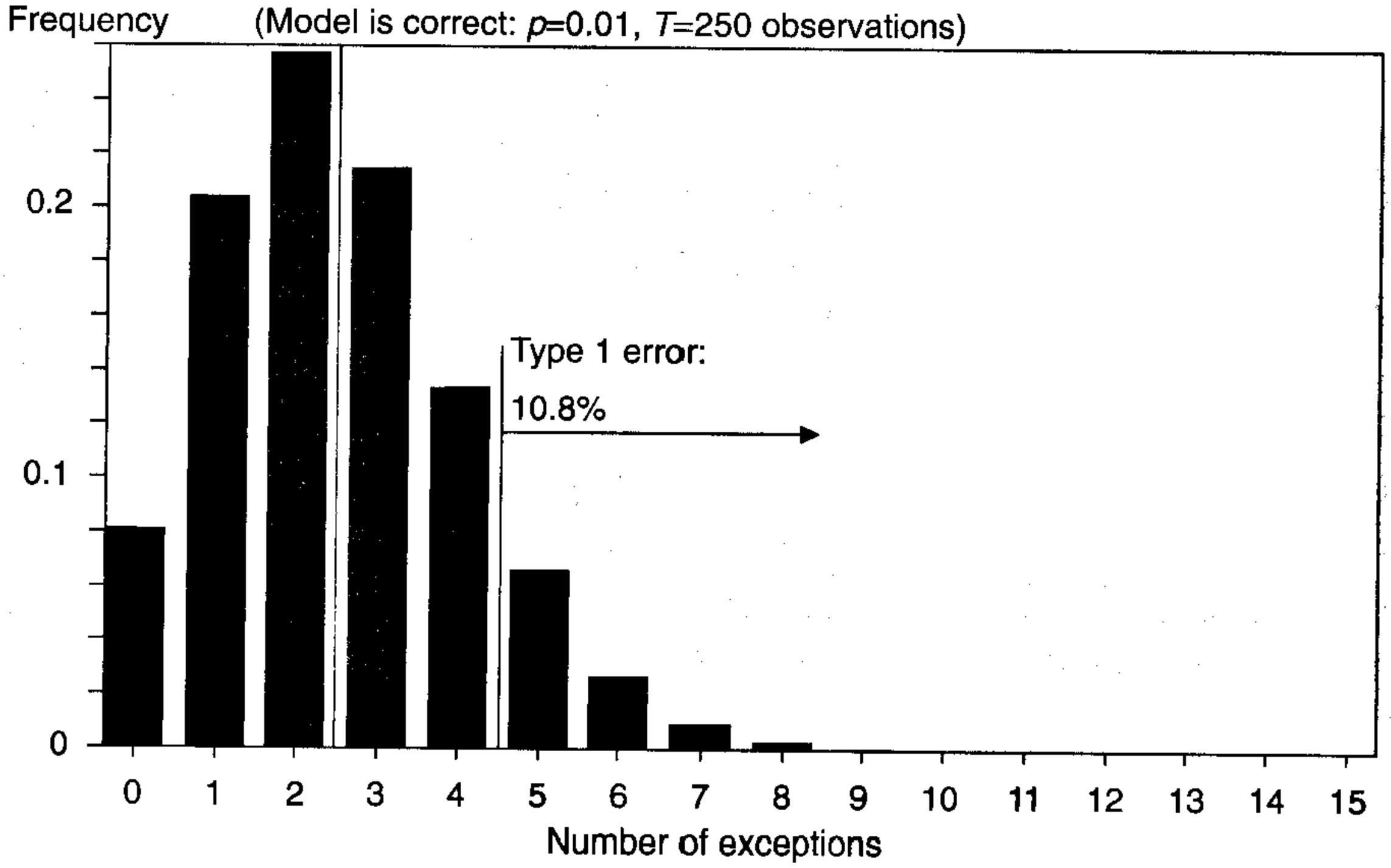
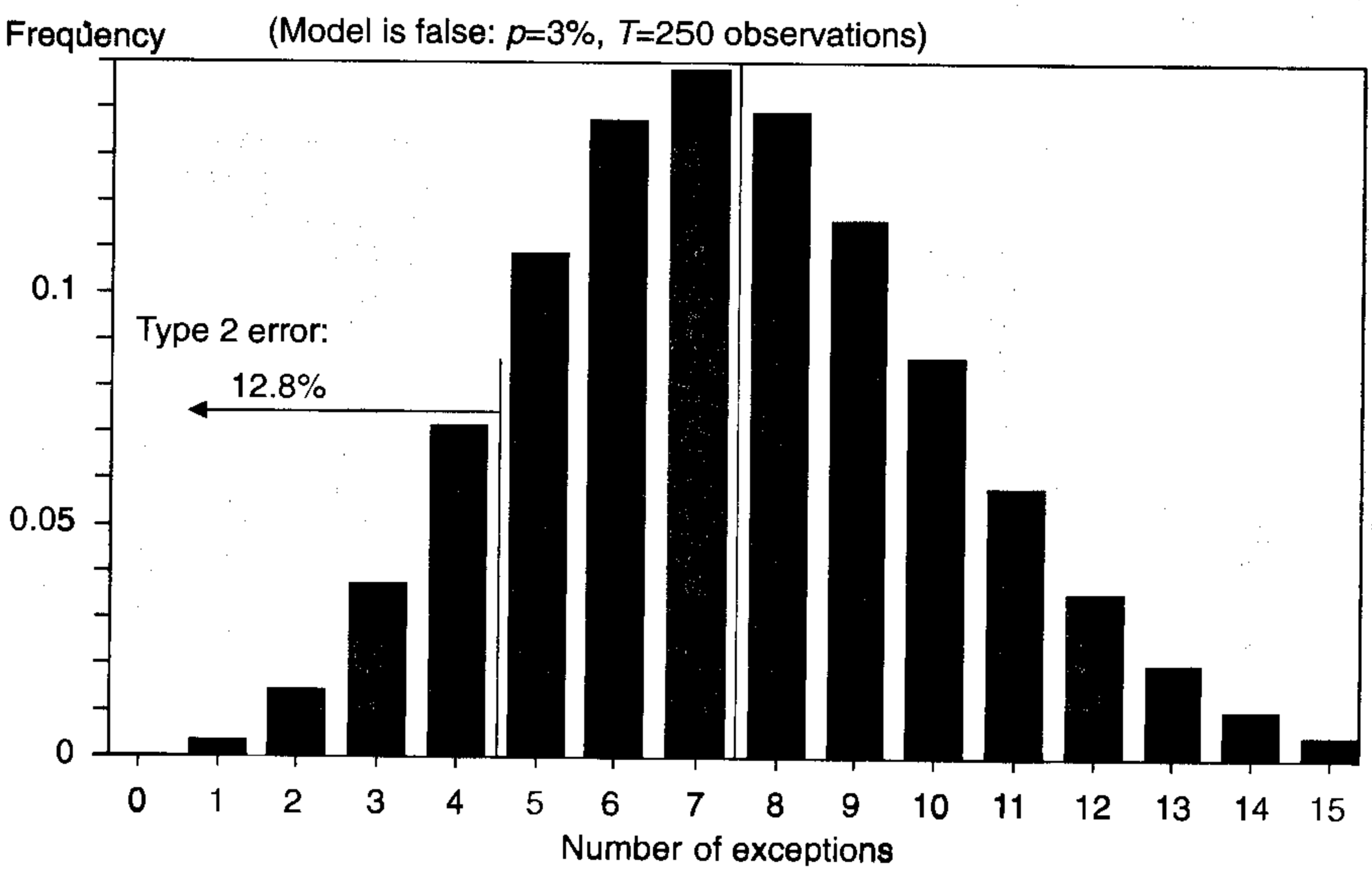


FIGURE 6-3

Distribution of exceptions when model is incorrect.



When designing a verification test, the user faces a tradeoff between these two types of error. Table 6-1 summarizes the two states of the world, correct versus incorrect model, and the decision. For backtesting purposes, users of VAR models need to balance type 1 errors against type 2 errors. Ideally, one would want to set a low type 1 error rate and then have a test that creates a very low type 2 error rate, in which case the test is said to be *powerful*. It should be noted that the choice of the confidence level for the decision rule is not related to the quantitative level p selected for VAR. This confidence level refers to the decision rule to reject the model.

Kupiec (1995) develops approximate 95 percent confidence regions for such a test, which are reported in Table 6-2. These regions are defined by the tail points of the log-likelihood ratio:

TABLE 6-1

Decision Errors

	Model	
Decision	**Correct**	**Incorrect**
Accept	OK	Type 2 error
Reject	Type 1 error	OK

TABLE 6-2

Model Backtesting, 95% Nonrejection Test Confidence Regions

		Nonrejection Region for Number of Failures N		
Probability level p	**VAR Confidence Level c**	**$T = 252$ Days**	**$T = 510$ Days**	**$T = 1000$ Days**
0.01	99%	$N < 7$	$1 < N < 11$	$4 < N < 17$
0.025	97.5%	$2 < N < 12$	$6 < N < 21$	$15 < N < 36$
0.05	95%	$6 < N < 20$	$16 < N < 36$	$37 < N < 65$
0.075	92.5%	$11 < N < 28$	$27 < N < 51$	$59 < N < 92$
0.10	90%	$16 < N < 36$	$38 < N < 65$	$81 < N < 120$

Note: N is the number of failures that could be observed in a sample size T without rejecting the null hypothesis that p is the correct probability at the 95 percent level of test confidence.
Source: Adapted from Kupiec (1995).

$$\mathrm{LR}_{uc} = -2\ln[(1-p)^{T-N}p^{N}] + 2\ln\{[1-(N/T)]^{T-N}(N/T)^{N}\} \quad (6.3)$$

which is asymptotically (i.e., when T is large) distributed chi-square with one degree of freedom under the null hypothesis that p is the true probability. Thus we would reject the null hypothesis if LR > 3.841. This test is equivalent to Equation (6.2) because a chi-square variable is the square of a normal variable.

In the JPM example, we had $N = 20$ exceptions over $T = 252$ days, using $p = 95$ percent VAR confidence level. Setting these numbers into Equation (6.3) gives $\mathrm{LR}_{uc} = 3.91$. Therefore, we reject unconditional coverage, as expected.

For instance, with 2 years of data ($T = 510$), we would expect to observe $N{=}pT{=}1$ percent times $510 = 5$ exceptions. But the VAR user will not be able to reject the null hypothesis as long as N is within the $[1 < N < 11]$ confidence interval. Values of N greater than or equal to 11 indicate that the VAR is too low or that the model understates the probability of large losses. Values of N less than or equal to 1 indicate that the VAR model is overly conservative.

The table also shows that this interval, expressed as a proportion N/T, shrinks as the sample size increases. Select, for instance, the $p = 0.05$ row. The interval for $T = 252$ is [6/252 = 0.024, 20/252 = 0.079]; for $T = 1000$, it is [37/1000=0.037, 65/1000 = 0.065]. Note how the interval shrinks as the sample size extends. With more data, we should be able to reject the model more easily if it is false.

The table, however, points to a disturbing fact. For small values of the VAR parameter p, it becomes increasingly difficult to confirm deviations. For instance, the nonrejection region under $p = 0.01$ and $T = 252$ is $[N < 7]$. Therefore, there is no way to tell if N is abnormally small or whether the model systematically overestimates risk. Intuitively, detection of systematic biases becomes increasingly difficult for low values of p because the exceptions in these cases are very rare events.

This explains why some banks prefer to choose a higher VAR confidence level, such as $c = 95$ percent, in order to be able to observe sufficient numbers of deviations to validate the model. A multiplicative factor then is applied to translate the VAR figure into a safe capital cushion number. Too often, however, the choice of the confidence level appears to be made without regard for the issue of VAR backtesting.

6.2.2 The Basel Rules

This section now turns to a detailed analysis of the Basel Committee rules for backtesting. While we can learn much from the Basel framework, it is important to recognize that regulators operate under different constraints from financial institutions. Since they do not have access to every component of the models, the approach is perforce implemented at a broader level. Regulators are also responsible for constructing rules that are comparable across institutions.

The Basel (1996a) rules for backtesting the internal-models approach are derived directly from this failure rate test. To design such a test, one has to choose first the type 1 error rate, which is the probability of rejecting the model when it is correct. When this happens, the bank simply suffers bad luck and should not be penalized unduly. Hence one should pick a test with a low type 1 error rate, say, 5 percent (depending on its cost). The heart of the conflict is that, inevitably, the supervisor also will commit type 2 errors for a bank that willfully cheats on its VAR reporting.

The current verification procedure consists of recording daily exceptions of the 99 percent VAR over the last year. One would expect, on average, 1 percent of 250, or 2.5 instances of exceptions over the last year.

The Basel Committee has decided that up to four exceptions are acceptable, which defines a "green light" zone for the bank. If the number of exceptions is five or more, the bank falls into a "yellow" or "red" zone and incurs a progressive penalty whereby the multiplicative factor k is increased from 3 to 4, as described in Table 6-3. An incursion into the "red" zone generates an automatic penalty.

TABLE 6-3

The Basel Penalty Zones

Zone	Number of Exceptions	Increase in k
Green	0 to 4	0.00
Yellow	5	0.40
	6	0.50
	7	0.65
	8	0.75
	9	0.85
Red	10+	1.00

Within the "yellow" zone, the penalty is up to the supervisor, depending on the reason for the exception. The Basel Committee uses the following categories:

- *Basic integrity of the model.* The deviation occurred because the positions were reported incorrectly or because of an error in the program code.
- *Model accuracy could be improved.* The deviation occurred because the model does not measure risk with enough precision (e.g., has too few maturity buckets).
- *Intraday trading.* Positions changed during the day.
- *Bad luck.* Markets were particularly volatile or correlations changed.

The description of the applicable penalty is suitably vague. When exceptions are due to the first two reasons, the penalty "should" apply. With the third reason, a penalty "should be considered." When the deviation is traced to the fourth reason, the Basel document gives no guidance except that these exceptions should "be expected to occur at least some of the time." These exceptions may be excluded if they are the "result of such occurrences as sudden abnormal changes in interest rates or exchange rates, major political events, or natural disasters." In other words, bank supervisors want to keep the flexibility to adjust the rules in turbulent times as they see fit.

The crux of the backtesting problem is separating back luck from a faulty model, or balancing type 1 errors against type 2 errors. Table 6-4 displays the probabilities of obtaining a given number of exceptions for a correct model (with 99 percent coverage) and incorrect model (with only 97 percent coverage). With five exceptions or more, the cumulative probability, or type 1 error rate, is 10.8 percent. This is rather high to start with. In the current framework, one bank out of 10 could be penalized even with a correct model.

Even worse, the type 2 error rate is also very high. Assuming a true 97 percent coverage, the supervisor will give passing grades to 12.8 percent of banks that have an incorrect model. The framework therefore is not very powerful. And this 99 versus 97 percent difference in VAR coverage is economically significant. Assuming a normal distribution, the true VAR would be 23.7 percent times greater than officially reported, which is substantial.

TABLE 6-4

Basel Rules for Backtesting, Probabilities of Obtaining Exceptions ($T = 250$)

		Coverage = 99% Model Is Correct		Coverage = 97% Model Is Incorrect		
Zone	Number of Exceptions N	Probability $P(X = N)$	Cumulative (Type 1) (Reject) $P(X \geq N)$	Probability $P(X = N)$	Cumulative (Type 2) (Do not reject) $P(X < N)$	Power (Reject) $P(X \geq N)$
Green	0	8.1	100.0	0.0	0.0	100.0
	1	20.5	91.9	0.4	0.0	100.0
	2	25.7	71.4	1.5	0.4	99.6
	3	21.5	45.7	3.8	1.9	98.1
Green	4	13.4	24.2	7.2	5.7	94.3
Yellow	5	6.7	10.8	10.9	12.8	87.2
	6	2.7	4.1	13.8	23.7	76.3
	7	1.0	1.4	14.9	37.5	62.5
	8	0.3	0.4	14.0	52.4	47.6
Yellow	9	0.1	0.1	11.6	66.3	33.7
Red	10	0.0	0.0	8.6	77.9	21.1
	11	0.0	0.0	5.8	86.6	13.4

The lack of power of this framework is due to the choice of the high VAR confidence level (99 percent) that generates too few exceptions for a reliable test. Consider instead the effect of a 95 percent VAR confidence level. (To ensure that the amount of capital is not affected, we could use a larger multiplier k.) We now have to decide on the cutoff number of exceptions to have a type 1 error rate similar to the Basel framework. With an average of 13 exceptions per year, we choose to reject the model if the number of exceptions exceeds 17, which corresponds to a type 1 error of 12.5 percent. Here we controlled the error rate so that it is close to the 10.8 percent for the Basel framework. But now the probability of a type 2 error is lower, at 7.4 percent only.[3] Thus, simply changing the VAR confidence level from 99 to 95 percent sharply reduces the probability of not catching an erroneous model.

[3] Assuming again a normal distribution and a true VAR that is 23.7 percent greater than the reported VAR, for an alternative coverage of 90.8 percent.

Another method to increase the power of the test would be to increase the number of observations. With $T = 1000$, for instance, we would choose a cutoff of 14 exceptions, for a type 1 error rate of 13.4 percent and a type 2 error rate of 0.03 percent, which is now very small. Increasing the number of observations drastically improves the test.

6.2.3 Conditional Coverage Models

So far the framework focuses on *unconditional coverage* because it ignores conditioning, or time variation in the data. The observed exceptions, however, could cluster or "bunch" closely in time, which also should invalidate the model.

With a 95 percent VAR confidence level, we would expect to have about 13 exceptions every year. In theory, these occurrences should be evenly spread over time. If, instead, we observed that 10 of these exceptions occurred over the last 2 weeks, this should raise a red flag. The market, for instance, could experience increased volatility that is not captured by VAR. Or traders could have moved into unusual positions or risk "holes." Whatever the explanation, a verification system should be designed to measure proper *conditional coverage*, that is, conditional on current conditions. Management then can take the appropriate action.

Such a test has been developed by Christoffersen (1998), who extends the LR_{uc} statistic to specify that the deviations must be serially independent. The test is set up as follows: Each day we set a deviation indicator to 0 if VAR is not exceeded and to 1 otherwise. We then define T_{ij} as the number of days in which state j occurred in one day while it was at i the previous day and π_i as the probability of observing an exception conditional on state i the previous day. Table 6-5 shows how to construct a table of conditional exceptions.

If today's occurrence of an exception is independent of what happened the previous day, the entries in the second and third columns should be identical. The relevant test statistic is

$$\begin{aligned} LR_{ind} = &- 2 \ln [(1 - \pi)^{(T_{00} + T_{10})} \pi^{(T_{01} + T_{11})}] + \\ &2 \ln [(1 - \pi_0)^{T_{00}} \pi_0^{T_{01}} (1 - \pi_1)^{T_{10}} \pi_1^{T_{11}}] \end{aligned} \tag{6.4}$$

Here, the first term represents the maximized likelihood under the hypothesis that exceptions are independent across days, or $\pi = \pi_0 = \pi_1 = (T_{01} + T_{11})/T$. The second term is the maximized likelihood for the observed data.

TABLE 6-5

Building an Exception Table: Expected Number of Exceptions

	Conditional		
	Day Before		
	No Exception	Exception	Unconditional
Current day			
No exception	$T_{00} = T_0 (1 - \pi_0)$	$T_{10} = T_1 (1 - \pi_1)$	$T(1 - \pi)$
Exception	$T_{01} = T_0 (\pi_0)$	$T_{11} = T_1 (\pi_1)$	$T(\pi)$
Total	T_0	T_1	$T = T_0 + T_1$

The combined test statistic for conditional coverage then is

$$\mathrm{LR}_{cc} = \mathrm{LR}_{uc} + \mathrm{LR}_{\mathrm{ind}} \tag{6.5}$$

Each component is independently distributed as $\chi^2(1)$ asymptotically. The sum is distributed as $\chi^2(2)$. Thus we would reject at the 95 percent test confidence level if LR > 5.991. We would reject independence alone if $\mathrm{LR}_{\mathrm{ind}} > 3.841$.

As an example, assume that JPM observed the following pattern of exceptions during 1998. Of 252 days, we have 20 exceptions, which is a fraction of $\pi = 7.9$ percent. Of these, 6 exceptions occurred following an exception the previous day. Alternatively, 14 exceptions occurred when there was none the previous day. This defines conditional probability ratios of $\pi_0 = 14/232 = 6.0$ percent and $\pi_1 = 6/20 = 30.0$ percent. We seem to have a much higher probability of having an exception following another one. Setting these numbers into Equation (6.4), we find $\mathrm{LR}_{\mathrm{ind}} = 9.53$. Because this is higher than the cutoff value of 3.84, we reject independence. Exceptions do seem to cluster abnormally. As a result, the risk manager may want to explore models that allow for time variation in risk, as developed in Chapter 9.

6.2.4 Extensions

We have seen that the standard exception tests often lack power, especially when the VAR confidence level is high and when the number of observations is low. This has led to a search for improved tests.

	Conditional		
	Day Before		
	No Exception	Exception	Unconditional
Current day			
No exception	218	14	232
Exception	14	6	20
Total	232	20	252

The problem, however, is that statistical decision theory has shown that this exception test is the most powerful among its class. More effective tests would have to focus on a different hypothesis or use more information.

For example, Crnkovic and Drachman (1996) developed a test focusing on the entire probability distribution, based on the *Kuiper statistic*. This test is still nonparametric but is more powerful. However, it uses other information than the VAR forecast at a given confidence level. Another approach is to focus on the time period between exceptions, called *duration*. Christoffersen and Pelletier (2004) show that duration-based tests can be more powerful than the standard test when risk is time-varying.

Finally, backtests could use parametric information instead. If the VAR is obtained from a multiple of the standard deviation, the risk manager could test the fit between the realized and forecast volatility. This would lead to more powerful tests because more information is used. Another useful avenue would be to backtest the portfolio components as well. From the viewpoint of the regulator, however, the only information provided is the daily VAR, which explains why exception tests are used most commonly nowadays.

6.3 APPLICATIONS

Berkowitz and O'Brien (2002) provide the first empirical study of the accuracy of internal VAR models, using data reported to U.S. regulators. They describe the distributions of P&L, which are compared with the VAR forecasts. Generally, the P&L distributions are symmetric, although they display fatter tails than the normal. Stahl et al. (2006) also report that, although the components of a trading portfolio could be strongly nonnormal, aggregation to the highest level of a bank typically produces symmetric distributions that resemble the normal.

FIGURE 6-4

Bank VAR and trading profits.

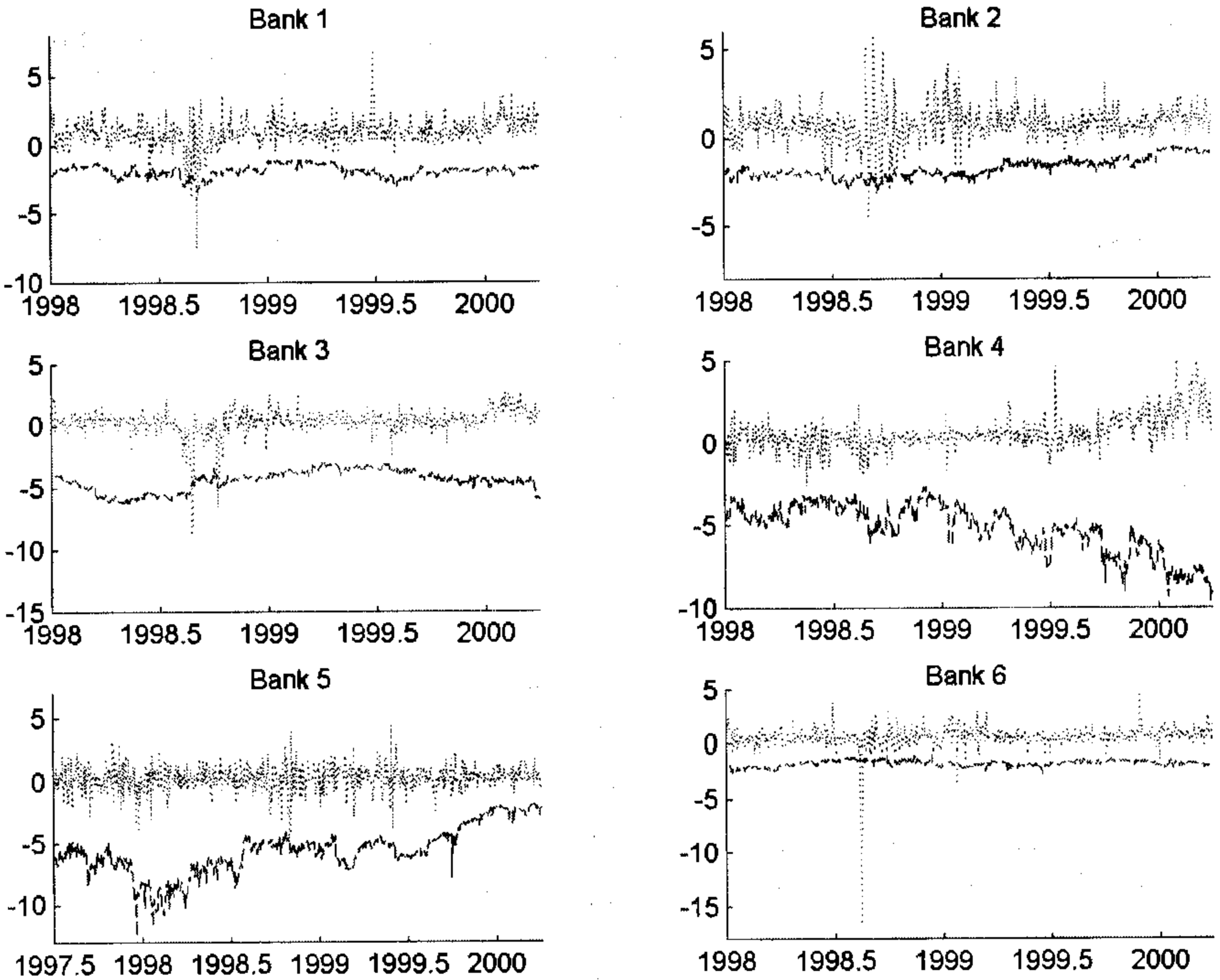


Figure 6-4 plots the time series of P&L along with the daily VAR (the lower lines) for a sample of six U.S. commercial banks. With approximately 600 observations, we should observe on average 6 violations, given a VAR confidence level of 99 percent.

It is striking to see the abnormally small number of exceptions, even though the sample includes the turbulent 1998 period. Bank 4, for example, has zero exceptions over this sample. Its VAR is several times greater than the magnitude of extreme fluctuations in its P&L. Indeed, for banks 3 to 6, the average VAR is at least 60 percent higher than the actual 99th percentile of the P&L distribution. Thus banks report VAR measures that are *conservative,* or too large relative to their actual risks. These results are surprising because they imply that the banks' VAR and hence their market-risk charges are too high. Banks therefore allocate too much regulatory capital to their trading activities. Box 6-2 describes a potential explanation, which is simplistic.

BOX 6-2

NO EXCEPTIONS

The CEO of a large bank receives a daily report of the bank's VAR and P&L. Whenever there is an exception, the CEO calls in the risk officer for an explanation.

Initially, the risk officer explained that a 99 percent VAR confidence level implies an average of 2 to 3 exceptions per year. The CEO is never quite satisfied, however. Later, tired of going "upstairs," the risk officer simply increases the confidence level to cut down on the number of exceptions.

Annual reports suggest that this is frequently the case. Financial institutions routinely produce plots of P&L that show no violation of their 99 percent confidence VAR over long periods, proclaiming that this supports their risk model.

Perhaps these observations could be explained by the use of actual instead of hypothetical returns.[4] Or maybe the models are too simple, for example failing to account for diversification effects. Yet another explanation is that capital requirements are currently not binding. The amount of economic capital U.S. banks currently hold is in excess of their regulatory capital. As a result, banks may prefer to report high VAR numbers to avoid the possibility of regulatory intrusion. Still, these practices impoverish the informational content of VAR numbers.

6.4 CONCLUSIONS

Model verification is an integral component of the risk management process. Backtesting VAR numbers provides valuable feedback to users about the accuracy of their models. The procedure also can be used to search for possible improvements.

Due thought should be given to the choice of VAR quantitative parameters for backtesting purposes. First, the horizon should be as short as possible in order to increase the number of observations and to mitigate

[4] Including fees increases the P&L, reducing the number of violations. Using hypothetical income, as currently prescribed in the European Union, could reduce this effect. Jaschke, Stahl, and Stehle (2003) compare the VARs for 13 German banks and find that VAR measures are, on average, less conservative than for U.S. banks. Even so, VAR forecasts are still too high.

the effect of changes in the portfolio composition. Second, the confidence level should not be too high because this decreases the effectiveness, or power, of the statistical tests.

Verification tests usually are based on "exception" counts, defined as the number of exceedences of the VAR measure. The goal is to check if this count is in line with the selected VAR confidence level. The method also can be modified to pick up bunching of deviations.

Backtesting involves balancing two types of errors: rejecting a correct model versus accepting an incorrect model. Ideally, one would want a framework that has very high power, or high probability of rejecting an incorrect model. The problem is that the power of exception-based tests is low. The current framework could be improved by choosing a lower VAR confidence level or by increasing the number of data observations.

Adding to these statistical difficulties, we have to recognize other practical problems. Trading portfolios do change over the horizon. Models do evolve over time as risk managers improve their risk modeling techniques. All this may cause further structural instability.

Despite all these issues, backtesting has become a central component of risk management systems. The methodology allows risk managers to improve their models constantly. Perhaps most important, backtesting should ensure that risk models do not go astray.

QUESTIONS

1. Define backtesting and exceptions.
2. Assume that a bank's backtests fail using the actual P&L return but not using the hypothetical return. Should the risk manager reexamine the risk model?
3. How is *type 1 error* different from *type 2 error* for a decision rule? Explain the meaning of these errors for backtesting the trading book of a bank. Can both errors be avoided?
4. For a fixed type 1 error rate, how can a test minimize the probability of a type 2 error?
5. Say that a bank reports 9 exceptions to its 99 percent daily VAR over the last year (252 days). Give two interpretations of this observation.
6. A bank reports 9 exceptions to its 99 percent VAR over the last year (252 days). Using the normal approximation to the binomial distribution, compute the z-statistic, and discuss whether the results would justify rejecting the model.

7. Backtesting is usually conducted on a short horizon, such as daily returns. Explain why.
8. A commercial bank subject to the Basel market-risk charge reports 4 exceptions over the last year. What is the multiplier k? Repeat with 10 exceptions.
9. Why is it useful to consider not only unconditional coverage but also conditional coverage?
10. A bank reports 6 exceptions to its 99 percent VAR over the last year (252 days), including 4 that follow another day of exception. Compute the likelihood-ratio tests, and discuss whether unconditional and conditional coverage is rejected.
11. The Berkowitz and O'Brien study indicates that bank are *conservative*, that is, generate VAR forecasts that are too large in relation to actual risks. What could explain this observation?