Model Validation Analytics in DROP

v4.05 11 February 2019

Probability Integral Transform

Introduction

1. Definition of Probability Integral Transform: In statistics, the probability integral transform or transformation relates to the results that data values that are modeled as being random variables from any continuous distribution can be converted to random variables having a standard uniform distribution (Dodge (2003), Wikipedia (2018)).
2. Exact vs. Approximate PIT Map: This holds exactly provided that the distribution being used is the true distribution of the random variables; if the distribution is the one fitted to the data, this will hold approximately in large samples.
3. Alternate Target PIT Distributions: The result is sometimes modified or extended so that the result of the transformation is a standard distribution other than the uniform distribution, such as the exponential distribution.

Applications

1. PIT for Statistical Hypothesis Testing: One use for the probability integral transform in statistical data analysis is to provide the basis for testing whether a set of observations can be reasonably modeled as arising from a specified distribution.

2. Using PIT to generate $U[0, 1]$: Specifically, the probability integral transform is applied to construct an equivalent set of values, and a test is then made to check whether a uniform distribution is appropriate for the constructed data set. Examples of this are the P-P plots and the Kolmogorov-Smirnov tests.
3. Copula Tests for Multivariate Data: A second use of the transformation is in the theory related to copulas which are a means of defining and working with distributions for statistically dependent multivariate data.
4. Applying PIT to Marginal Distributions: Here the problem of defining or manipulating a joint probability distribution for a set of random variables is simplified or reduced in apparent complexity by applying the probability integral transform to each of the components and then working with a joint distribution for which the marginal variables have uniform distributions.
5. PIT for Inverse Transform Sampling: A third use is based on applying then inverse of the probability integral transform to convert random variables from a uniform distribution to have the selected distribution; this is known as inverse transform sampling.

Statement and Proof

1. Statement: Suppose a random variable X has a continuous distribution for which the cumulative distribution function (CDF) is F_X. Then the random variable Y is defined as

$$Y = F_X(X)$$

   It has a uniform distribution.
2. Proof: Given any random continuous variable X, define

$$Y = F_X(X)$$

Then

$$\begin{aligned} F_Y(y) &= \mathbb{P}[Y \leq y] \\ &= \mathbb{P}[F_X(X) \leq y] \\ &= \mathbb{P}\left[X \leq {F_X}^{-1}(y)\right] \\ &= F_X\left({F_X}^{-1}(y)\right) \\ &= y \end{aligned}$$

If F_Y is the CDF of a uniform $[0, 1]$ random variable, Y will have a uniform distribution on the interval $[0, 1]$.

Examples

1. Underlying Univariate Distribution – Standard Normal: As an illustrative example, let X be a random variable with the Standard Normal distribution $\mathcal{N}(0, 1)$. Then its CDF is

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \quad x \in \mathbb{R}$$

   where $\text{erf}(\cdot)$ is the error function. Then the new random variable Y defined as

$$Y = \Phi(x)$$

   is uniformly distributed.
2. Underlying Univariate Distribution - Standard Exponential: If X has an exponential distribution with unit mean, then its CDF is

$$F(x) = 1 - e^{-x}$$

and the immediate result of the probability integral transform is that

$$Y = 1 - e^{-X}$$

has a uniform distribution. The symmetry of the uniform distribution can then be used to show that

$$Y' = e^{-X}$$

also has a uniform distribution.

References

- Dodge, Y. (2003): *The Oxford Dictionary of Statistical Terms* **Oxford University Press**
- Wikipedia (2018): Probability Integral Transform

t-Statistic

Overview

1. Definition of the t-statistic Metric: In statistics, the **t-statistic** is the ratio of the departure of the estimated value of a parameter from its hypothesized value to its standardized error. It is used for hypothesis testing via the Student's t-test.
2. Population Mean from a Sample Distribution: For example, it is used in estimating the population mean from a sample distribution of sample means if the population standard deviation is unknown.

Definition and Features

1. Mathematical Expression for the t-statistic: Let $\hat{\beta}$ be an estimate of the parameter β in some statistical model. Then t-statistic for this parameter is any quantity of the form

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{s.e.(\hat{\beta})}$$

where β_0 is a non-random, known constant which may or may not match the actual unknown parameter value β, and $s.e.(\hat{\beta})$ is the standard error of the estimator $\hat{\beta}$ for β (Wikipedia (2018)).

2. t-statistic Reports from Statistical Packages: By default, statistical packages report t-statistic with

$$\beta_0 = 0$$

these t-statistics are used to test the significance of the corresponding regressor. However, when t-statistic is needed to test the hypothesis of the form

$$H_0 : \beta = \beta_0$$

then a non-zero β_0 may be used.

3. t-statistic Distribution for Linear Regression: If $\hat{\beta}$ is an ordinary least squares estimate in the classical linear regression model – that is with normally distributed and homoscedastic error terms – and if the true value of the parameters β is equal to β_0, then the sampling distribution of the t-statistic is the Student's t-distribution with $n - k$ degrees of freedom, where n is the number of observations, and k is the number of regressors, including the intercept.
4. Asymptotically Normal $\hat{\beta}$ and $s.e.(\hat{\beta})$: In the majority of models, the estimator $\hat{\beta}$ is consistent for β and is distributed asymptotically normally. If the true value of the parameter β is equal to β_0 and the quantity $s.e.(\hat{\beta})$ correctly estimates the asymptotic variance of the estimator, then the t-statistic will asymptotically have the standard normal distribution.
5. Examples of Asymptotically Non-normal t-statistic: In some models, the distribution of the t-statistic is different from the normal distribution, even asymptotically. For example, when the time series with unit root is regressed in the augmented Dickey-Fuller test, the test t-statistic will asymptotically have one of the Dickey-Fuller distributions, depending on the test setting.

Use

1. Common Use of the t-statistic: Most frequently, t-statistics are used in the Student's t-tests, a form of statistical hypothesis testing, and in the computation of certain confidence intervals.
2. t-statistic as a Pivotal Quantity: The key property of the t-statistic is that it is a pivotal quantity – while defined in terms of the sample means, its sampling distribution does not depend on the population parameters, and thus it can be used regardless of what these may be.
3. Difference between t-statistic and z-score: One can also divide the residual by the sample standard deviation:

$$g(x, X) = \frac{x - \bar{X}}{s}$$

to compute an estimate for the number of standard deviations a given sample is from the mean, as a sample version of the z-score, the z-score requiring the population parameters.

Prediction

1. As an Ancillary Pivotal Statistic: Given a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with unknown mean and variance, the t-statistic of a future observation X_{n+1} after one has made n observations, is an ancillary statistic – it is both a pivotal quantity since it does not depend on the values of μ and σ^2, as well as a statistic, since it may be computed from observations.
2. Predictive Confidence Interval as a t-distribution: This allows one to compute a frequentist prediction interval – a predictive confidence interval – via the following t-distribution

$$\frac{X_{n+1} - \overline{X_n}}{s_n\sqrt{1 + \frac{1}{n}}} \sim T^{n-1}$$

3. Estimating the Subsequent Confidence Interval: Solving for X_{n+1} yields the predictive distribution $\overline{X_n} + s_n\sqrt{1+\frac{1}{n}} \cdot T^{n-1}$ from which one may compute predictive confidence intervals – given a probability p, one may compute intervals such that $100p\%$ of the time, the next X_{n+1} will fall within that interval.

Related Concepts

1. z-score (Standardization): If the population of the parameters are known, then rather than computing the t-statistic, one may compute the z-score analogously, and rather than using t-test, one uses a z-test. This is rare outside of standardized testing.
2. Standardized Residual: In regression analysis, the standard errors of residuals at different data point vary, e.g., the comparison of the middle versus the end points of the standard linear regression shows this, and thus one must divide the different residuals for different estimates of the error, yielding what are called standardized residuals.

References

- Wikipedia (2018): t-statistic

p-value

Overview

1. p-value/Probability Value/Asymptotic Significance: In statistical model hypothesis testing, the ***p-value*** or ***probability value*** or ***asymptotic significance*** is the probability of a given statistical model that, under the assumption that the NULL hypothesis is true, the statistical summary – such as sample mean difference across two compared groups – would be greater than or equal to the actual observed results (Wasserstein and Lazar (2016), Wikipedia (2019)).
2. p-values in Statistical Hypothesis Testing: The use of p-values in statistical hypothesis testing is common in many fields of research (Bhattacharya and Habtzghi (2002)) such as economic, finance, political science, psychology (Wetzels, Matzke, Lee, Rouder, Iverson, and Wagenmakers (2011)), biology, criminal justice, criminology, and sociology (Babbie (2007)). Their misuse has been a matter of considerable controversy.
3. Representation of the Term in Literature: Italicization, capitalization, and hyphenation of the term varies. For example, AMA style uses *P value*, APA style uses *p value*, and the American Statistical Association uses *p-value* (American Statistical Association (2008)).

Basic Concepts

1. Definition of Statistical Hypothesis: In statistics, every conjecture concerning the unknown distribution F of a random variable X is called a *statistical hypothesis*.
2. Definition of Statistical Tests: Methods of verifying statistical hypothesis are called statistical tests.
3. Definition of a Significance Test: If only a single hypothesis is stated and the aim is to verify if this hypothesis is true, but not at the same time to investigate the other hypothesis, such a test is then called a *significance test*.
4. Definition of Parametric Hypothesis: A statistical hypothesis that refers only to the numerical values of the unknown distribution is called a *parametric hypothesis*.
5. Definition of Parametric Tests: Tests of parametric hypothesis are called *parametric tests*.
6. Definitions of Non-parametric Hypotheses and Tests: One can likewise have *non-parametric hypotheses* and *non-parametric tests*.
7. Reductio ad Absurdum Argument adapted to Statistics: The p-value is used in the context of a NULL hypothesis testing in order to quantify the idea of statistical significance of the evidence – note that the statistical significance of a result does not imply that the result is significant scientifically as well. NULL hypothesis testing is a *reductio ad absurdum* argument adapted to statistics. In essence, a claim is valid if its counter-claim is improbable.
8. Definition of the NULL Hypothesis: As such, the only hypothesis that needs to be specified in this test and that which embodies the counter-claim is referred to as the *NULL Hypothesis*.
9. Definition of Statistical Significance: A result is said to be *statistically significant* if it allows the rejection of the NULL hypothesis. That is, as per the reductio ad absurdum reasoning, the statistically significant should be highly improbable if the NULL hypothesis is assumed to be true.
10. Acceptable Alternatives to the NULL Hypothesis: The rejection of the NULL hypothesis implies that the correct hypothesis lies in the logical complement of the NULL hypothesis. However, unless there is a single alternative to the NULL hypothesis, the rejection of the NULL hypothesis does not tell which of the alternatives is the correct one.
11. Example - Standard Normal NULL Hypothesis: As a general example, if a NULL hypothesis is assumed to follow the standard normal distribution $\mathcal{N}(0,1)$ then the rejection of this NULL hypothesis can either mean:

a. The mean is not zero, OR
b. The variance is not unity, OR
c. The distribution is not normal, depending on the type of test performed.

12. Rejecting Standard Normal NULL Hypothesis: However, supposing one manages to reject the mean zero hypothesis, even if one knows that the distribution is normal and that the variance is unity, the NULL hypothesis does not tell which non-zero value should be adopted as the new mean.
13. Observation Occurrence under the NULL Hypothesis: If X is a random variable representing the observed data and H is the statistical hypothesis under consideration, then the notion of statistical significance can be naively quantified by the conditional probability $\mathbb{P}[A|H]$, which gives the likelihood of certain observation event A if the hypothesis is *assumed* to be correct.
14. Inapplicability under Continuous Random Variable: However, if X is a continuous random variable, the probability of observing a specific instance x is zero, so that

$$\mathbb{P}[X = x|H] = 0$$

Thus, the naïve definition is inadequate and needs to be changed so as to accommodate the continuous random variables.

15. Misinterpretation as Continuous Bayesian Probabilities: Nevertheless, it helps to clarify that the p-values should *not* be confused with the probability of the hypothesis (as is done in Bayesian hypothesis testing) such as $\mathbb{P}[H|A]$, the probability of the hypothesis given the data, or $\mathbb{P}[H]$, the probability of the hypothesis being true, or $\mathbb{P}[A]$, the probability of observing the data.

Definition and Interpretation

1. Revisiting the Definition of p-value: The p-value is defined as the probability, under the NULL hypothesis, of obtaining a result equal to, or more extreme than, what was actually observed. It is often denoted as H_0, as opposed to H_a, which is sometimes used to represent the alternate hypothesis.
2. Left-Tailed and Right-Tailed Events: Depending on how it is looked at, the *more extreme than what was actually observed* can mean $\{X \geq x\}$ - a right-tailed event, $\{X \leq x\}$ - a left-tailed event, or the ***smaller*** of $\{X \leq x\}$ and $\{X \geq x\}$ - a double-tailed event.
3. Mathematical Specification of the p-value: Thus, the p-value is given by $\mathbb{P}[X \geq x|\ H]$ for the right-tailed event, $\mathbb{P}[X \leq x|H]$ for the left-tailed event, and $2\min(\mathbb{P}[X \geq x|\ H], \mathbb{P}[X \leq x|H])$ for the double-tailed event.
4. Higher Significance under Smaller p-value: The smaller the p-value, the higher the significance, because it tells the investigator that the hypothesis under consideration may not adequately explain the observation.
5. Pre-defined Threshold of Significance α: The NULL hypothesis H is rejected if any of these probabilities is less than or equal to a small, fixed but arbitrarily pre-defined threshold value α, which is referred to as the *level of significance*.
6. Characteristics of and Typical Choices of α: Unlike the p-value, the α level is not derived from any observational data, and does not depend on the underlying hypothesis; the value of α is instead set by the researcher before examining the data. The setting of α is arbitrary. By convention, α is commonly set to 0.05, 0.01, 0.005, or 0.001.
7. Randomness of the Resulting p-value: Since the value of x that defines the left-tail or the right-tail is a random variable, this makes the p-value a function of x and a random variable in itself; under the NULL hypothesis, the p-value is defined uniformly over the $[0, 1]$ interval, assuming x is continuous. Thus, the p-value is not fixed.
8. Variation in p-values across Observations: This implies that the p-value cannot be given a frequency counting interpretation since the probability has to be fixed for the frequency counting interpretation to hold. In other words, if the same test is repeatedly independently bearing upon the same overall NULL hypothesis, it will yield different p-values at every repetition. Nevertheless, the different p-values can be combined using Fischer's combined probability test.

9. Interpreting the Generated p-value Set: It should further be noted that an *instantiation* of this p-value can be given a frequency counting interpretation with respect to the number of observations taken during a given test, as per the definition, as the percentage of observations more extreme than then one observed under the assumption that the NULL hypothesis is true.

Misconceptions

1. Misuse and Misinterpretation of the p-value: There is widespread agreement that the p-values are often misused and misinterpreted (Goodman (1999), Scientific American (2015), Wasserstein and Lazar (2016)). One practice that has been particularly criticized is accepting the alternative hypothesis for any p-value nominally less than 0.05 without other supporting evidence.
2. Factors deciding the p-value Outcome: Although p-values are helpful in assessing how incompatible the data are with a specified statistical model, contextual factors must also be considered, such as the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of the assumptions that underlie the data analysis (Wasserstein and Lazar (2016)).
3. p-value as a NULL Hypothesis Probability: Another concern is that the p-value is often misunderstood as being the probability that the NULL hypothesis is true (Colquhoun (2014), Wasserstein and Lazar (2016)).
4. Alternative Measures of Hypothesis Verification: Some statisticians have proposed replacing p-values with alternate measures of evidence (Wasserstein and Lazar (2016)) such as confidence levels (Ranstam (2012), Lee (2017)), likelihood ratios (Perneger (2001), Royall (2004)), or Bayes' factors (Marden (2000), Schimmack (2015), Stern (2016)), but there is heated debate on the feasibility of these alternatives (Murtaugh (2014), Aschwanden (2016)).
5. p-values as Continuous Index of Strength: Others have suggested removing the fixed significance thresholds and interpreting p-values as the continuous index of the strength of

evidence against the NULL hypothesis (Amrhein, Korner-Nievergelt, and Roth (2017), Amrhein and Greenland (2017)).

6. p-values alongside Prior Probabilities: Yet, others have suggested reporting alongside p-values the prior probabilities of a real effect that would be required to obtain a false positive risk, i.e., the probability that there is no real effect, below a pre-specified threshold (Colquhoun (2017)).

Usage

1. p-value Usage and Experimental Design: The p-value is widely used in statistical hypothesis testing, specifically in NULL hypothesis significance testing., In this method, as part of the experimental design, before performing the experiment, one first chooses the model – the NULL hypothesis – and a threshold value for p, called the significance level of the test, traditionally 0.05 or 0.01 (Nuzzo (2014)), and denoted as α.
2. p-value Based Data vs. Hypothesis Consistency: If the p-value is less than the chosen significance level α, that suggests that the observed data is significantly inconsistent with the NULL hypothesis, and that the NULL hypothesis may be rejected. However, that does not prove that the alternate hypothesis is true.
3. Type I Error Rate Limit: When the p-value is calculated correctly, this test guarantees that the type I error rate is about α. For typical analysis using the standard

$$\alpha = 0.05$$

cutoff, the NULL hypothesis is rejected when

$$p < 0.05$$

and not rejected when

$$p > 0.05$$

The p-value does not, in itself, support the reasoning about the probabilities of the hypothesis, but is only a tool for deciding whether to reject the NULL hypothesis.

Calculation

1. Definition of the Test Statistic: Usually, X is a test statistic, rather than any of the actual observations. The test statistic is an output of the scalar function of all observations.
2. Test Statistic as an Observation Summary: This statistic provides a single number, such as the average or the correlation coefficient, that summarizes the characteristics of the data, in a way relevant to a particular inquiry.
3. Distribution of the Test Statistic: As such, the test statistic follows the distribution provided by the function used to define that test statistic and the distribution of the input observational data.
4. Test Statistic under Normal Distribution: For the important case under which the data is hypothesized to follow the normal distribution, depending on the nature of the test statistic – and therefore the underlying hypothesis of the test statistic – different NULL hypotheses have been developed. Some such tests are the z-test for the normal distribution, t-test for Student's t-distribution, and f-test for the f-distribution.
5. Test-Statistic under Non-normal Distribution: When data do not follow a normal distribution, it may still be possible to approximate the distribution of the test statistic by a normal distribution by invoking the Central Limit Theorem for large samples, as is the case with Pearson's chi-squared test.

6. Components Needed for p-value Estimation: Thus, computing the p-value requires a NULL hypothesis, a test statistic, a determination of whether the researcher is performing a one-tailed or a two-tailed test, and the data.
7. Estimation of the Sampling Distribution: Even though computing the test statistic on a given data may be easy, computing the sampling distribution under the NULL hypothesis, and then computing its cumulative distribution function CDF, is often a difficult problem.
8. CDF to p-value Translation Tables: Today, this computation is done using statistical software, often via numerical methods rather than exact formula, but in the early and the mid-20th century, this was done instead via tables of values, and one interpolated or extrapolated p-values from these discrete values.
9. Quantile Inversions from Cumulative Probabilities: Rather than using a table of p-values, Fisher instead converted the CDF, publishing a list if test statistic for a given fixed set of p-values; this corresponds to computing the quantile function (or inverse CDF).

Distribution

1. p-value Distribution under NULL Hypothesis: When the NULL hypothesis is true, assuming that it takes the form

$$H_0: \theta = \theta_0$$

   and that the underlying random variable is continuous, the probability distribution of the p-value is uniform in the interval $[0, 1]$.
2. p-value Distribution under Alternate Hypothesis: By contrast, if the alternate hypothesis is true, the distribution is dependent on the sample size and the true value of the parameter being studied (Hung, O'Neill, Bauer, and Kohne (1997), Bhattacharya and Habtzghi (2002)).

3. Constructing p-Curve and Estimating p-Hacking: The distribution of p-values for a group of studies is called a p-curve (Head, Holman, Lanfear, Kahn, and Jennions (2015)). This curve is affected by four factors; the proportion of studies that examined the false NULL hypothesis, the power of the studies that investigated the false NULL hypothesis, the alpha levels, and the publication bias (Lakens (2015)). A p-Curve can be used to assess the reliability of the scientific literature, such as by detecting publication bias or p-hacking (Simonsohn, Nelson, and Simmons (2014), Head, Holman, Lanfear, Kahn, and Jennions (2015)).

Example – One Roll of a Pair of Dice

1. NULL Hypothesis – Unbiased Dice Outcome: Suppose a researcher rolls a pair of dice once and assumes the NULL hypothesis that then dice are fair and not loaded or weighted toward any specific number/roll/result, i.e., it is uniform.
2. Test Statistic of the Experiment: The test statistic is the *sum of the rolled numbers* and is one-tailed.
3. Outcome p-value vs. Acceptance Threshold: The researcher rolls the dice and observes that both dice show a 6, yielding a test statistic of 12. The p-value of the outcome is $\frac{1}{36}$ – because under the assumption of the NULL hypothesis, the test statistic is uniformly distributed, or about 0.028 – the highest test statistic out of

$$6 \times 6 = 36$$

possible outcomes. If the researcher assumed a significance level of 0.05, this result would be deemed significant and the hypothesis that the dice are fair would be rejected.

4. Importance of the Experimental Design: In this case, the single roll provides a very weak basis – that is insufficient data – to draw meaningful conclusion about the dice. This illustrates the danger with blindly applying p-values without considering the experimental design.

Example: Five Heads in a Row

1. NULL Hypothesis of Fair Coin: Suppose a researcher flips five coins in a row and assumes a NULL hypothesis that the coin is fair.
2. One/Two Tailed Test Statistic: The test statistic of *total number of heads* can be one-tailed or two-tailed; one-tailed corresponds seeing if the coin is biased toward heads, and a two-tailed test corresponds to seeing if the coin is biased either way.
3. Realized One-Tailed Test Statistic: The researcher flips the coin 5 times and observes head each time ($HHHHH$) yielding a test statistic of 5. In a one-tailed test this is the upper extreme of all possible outcomes, and yields a p-value of

$$\left(\frac{1}{2}\right)^5 = \frac{1}{32} = 0.03$$

4. Statistical Significance of the Test-Statistic: If the researcher assumed a significance level of 0.05, this result would be deemed significant and the hypothesis that the coin is fair would be rejected. In a two-tailed test, a test statistic of zero heads ($TTTTT$) is just as extreme, and thus the data of $HHHHH$ would yield a p-value of

$$2 \times \left(\frac{1}{2}\right)^5 = \frac{1}{16} \cong 0.06$$

which is not significant at the 0.05 level.

5. Impact of the Test Statistic Directionality: This demonstrates that specifying the direction on a symmetric test static halves the p-value, i.e., increases the significance, and can mean the difference between the data being considered significant or not.

Sample Size Dependence

1. Test Design and Tail Setting: Suppose a researcher flips a coin some arbitrary number of times n and assumes a NULL hypothesis that the coin is fair. The test statistic is the number of heads and the test itself is two-tailed.
2. p-values across Different Sample Sizes: Suppose the researcher observes heads for each flip, yielding a test statistic of n and a p-value of $\frac{2}{2^n}$. If the coin was flipped only 5 times, the p-value would be

$$\frac{2}{32} = 0.,0625$$

which is not significant at the 0.05 level. But if the coin was flipped 10 times, the p-value would be

$$\frac{2}{1024} \cong 0.002$$

which is significant at the 0.05 level.

3. Impact of the Sample Size: In the second case, the data suggest that the NULL hypothesis is false – that is, the coin is not fair somehow – but changing the sample size changes the p-

value. In the first case, the sample size is not large enough for the NULL hypothesis to be rejected at the 0.05 level – in fact, the p-value can never be 0.05 in this instance.

4. Sample Size as a Hypothesis Test Parameter: This demonstrates that, in interpreting p-values, one must also know the sample size, which complicates the analysis.

Alternating Coin Flips

1. Experimental Design and Test Statistic: Suppose a researcher flips a coin 10 times and assumes a NULL hypothesis that the coin is fair. The test-statistic is the total number of heads and is two tailed.
2. Experimental Outcome and Test Statistic: Suppose that the researcher observes alternating heads and tails with every flip ($HTHTHTHTHT$). This yields a test statistic of 5.
3. Alternate Test Statistic - Flip Count: Suppose that the test statistic for the experiment was instead the *number of alterations* – that is the number of times H follows T or T follows H – which is one-tailed.
4. Realization of the Test Statistic: That would yield a test statistic of 9, which is extreme and has a p-value of

$$\frac{2}{2^9} = \frac{1}{256} \cong 0.0039$$

That would be considered extremely significant, well beyond the 0.05 level.

5. Appropriateness of the Test Statistic: These data indicate that, in terms of one test statistic, the data set is entirely unlikely to have occurred by chance, but it does not indicate whether the coin is biased toward head or tail.
6. Contradictory Outcomes from the p-value Tests: By the first test statistic, the data yield a high p-value, suggesting that the number of heads is not very unlikely. By the second test

statistic, the data yield a low p-value, suggesting that the pattern of flips is very, very unlikely.

7. Possible Causes for the Contradiction: There is no *alternative hypothesis* – so only a rejection of the NULL hypothesis is possible – and such data could have many causes. For instance, the data may have been forged, or the coins may have been flipped by a magician who intentionally fabricated the outcomes.
8. Limitations Inherent in the p-value Approach: This example demonstrates that the p-value depends completely on the test statistic used and illustrates that the p-values can only help researchers reject a NULL hypothesis, not consider the other hypothesis.

Generic Coin Flipping

1. Use of a Different Test-Statistic: As an example of a different statistical test, an experiment is performed to determine whether a coin flip is fair, i.e., equal chance of landing heads or tails, or unfairly biased – one outcome being more likely than the other.
2. Experimental Design and Test Statistic: Suppose that the experimental results show the coin turning up heads 14 times out of a total of 20 flips. The NULL hypothesis is that the coin is fair, and the test statistic is the number of heads.
3. Estimating the Hypothesis' Right-Tail Probability: If a right-tailed test is considered, the p-value of this result is the chance of a fair coin landing on heads *at least* 14 out of 20 flips. That probability can be computed from the binomial coefficients as

$$\mathbb{P}[14\ Heads] + \mathbb{P}[15\ Heads] + \cdots + \mathbb{P}[20\ Heads] = \frac{1}{2^{20}}[C_{14}^{20} + C_{15}^{20} + \cdots + C_{20}^{20}]$$

$$= \frac{60460}{1048576} \cong 0.058$$

4. Computing the Two-Tailed p-value: This probability is the p-value, considering only the extreme results that favor heads. This is called a one-tailed test. However, the deviation can be in either direction, favoring heads or tails. The two-tailed p-value, which considers deviations favoring either heads or tails, may instead be calculated.
5. Applying the Fair Coin Hypothesis Symmetry: As the binomial distribution is symmetrical for a fair coin, the two-sided p-value is simply twice the above calculated single-sided p-value, that is 0.115.
6. Hypothesis Test and Outcome Summary: Therefore, in the above example:
    a. NULL Hypothesis (H_0): The coin is fair, with

$$\mathbb{P}[H] = 0.50$$

    b. Test Statistic: Number of Heads
    c. Level of Significance: 0.05
    d. Observation O: 14 out of 20 heads
    e. Two-tailed p-value of Observation O given H_0:

$$2\min(\mathbb{P}[Number\ of\ Heads \geq 14], \mathbb{P}[Number\ of\ Heads \leq 14]) = 2\min(0.058, 0.978) = 2 \times 0.058 = 0.115$$

7. Probability of the Left-Tailed Event: Note that

$$\mathbb{P}[Number\ of\ Heads \leq 14] = 1 - \mathbb{P}[Number\ of\ Heads \geq 14] = 1 - 0.058 + 0.036 = 0.978$$

   However, the symmetry of the binomial distribution makes it an unnecessary computation to find the smaller of the two probabilities.
8. p-value Clearing the Acceptance Threshold: Here the calculated p-value exceeds 0.05, so the observation is consistent with the NULL hypothesis, as it falls within the range of what would happen 95% of the time where the coin in fact fair. Hence the NULL hypothesis at the

5% level is not rejected. Although the coin did not fall evenly, the deviation from the expected outcome is small enough to be consistent with chance.

9. p-value Test under 15 Head Outcomes: However, had one more head been obtained, the resulting two-tailed p-value would have been 0.0414 (4.14%). The NULL hypothesis would be rejected when a 0.05 cutoff is used.

References

- American Statistician (2008): ASA House Style
- Amrhein, V., F. Korner-Nievergelt, and T. Roth (2017): The Earth is Flat (p > 0.05); Significance Thresholds and the Crisis of Unreplicable Research *PeerJ* **5** e3544
- Amrhein, V., and S. Greenland (2017): Remove, rather than Re-define, Statistical Significance
- Aschwanden, C. (2016): Statisticians Found One Thing They can Agree on: It's Time to Stop Misusing p-values
- Babbie, E. (2007): *The Practice of Social Research 11th Edition* **Thomson Wadsworth** Belmont CA
- Bhattacharya, B., and D. Habtzghi (2002): Median of the p-value under the Alternate Hypothesis *American Statistician* **56 (3)** 202-206
- Colquhoun, D. (2014): An Investigation of the False Discovery Rate and the Misinterpretation of the p-values *Royal Society Open Science* **1 (3)** 140216-140231
- Colquhoun, D. (2017): p-values *Royal Society Open Science* **4 (12)** 171085-171106
- Goodman, S. N. (1999): Toward Evidence-Based Medical Statistics; 1. The p-value Fallacy *Annals of Internal Medicine* **130 (12)** 995-1004
- Head, M. L., L. Holman, R, Lanfear, A. T. Kahn, and M. D. Jennions (2015): The Extent and Consequences of p-Hacking in Science *PLoS Biology* **13 (3)** e1002106

- Hung, H. M. J., R. T. O'Neill, P. Bauer, and K. Kohne (1997): The Behavior of the p-value when the Alternate Hypothesis is True *Biometrics* **53 (1)** 11-22
- Lakens, D. (2015): What p-Hacking really looks like: A Comment on Masicampo and LaLande (2012) *Quantitative Journal of Experimental Psychology (Hove)* **68 (4)** 829-832
- Lee, D. K. (2017): Alternatives to p-value: Confidence Interval and Effect Size *Korean Journal of Anesthesiology* **69 (6)** 555-562
- Marden, J. I. (2000): Hypothesis Testing: From p-values to Bayes' Factors *Journal of the American Statistical Association* **95 (452)** 1316-1320
- Murtaugh, P. A. (2014): In Defense of p-values *Ecology* **95 (3)** 611-617
- Nuzzo, R. (2014): Scientific Method, Statistical Errors *Nature* **506 (7487)** 150-152
- Perneger, T. V. (2001): Sifting the Evidence: Likelihood Ratios are Alternatives to p-values *British Medical Journal* **322 (7295)** 1184-1185
- Ranstam (2012): What the p-value Culture is Bad and Confidence Intervals are a better Alternative *Osteoarthritis and Cartilage* **20 (8)** 805-808
- Royall, R. (2004): *Likelihood Paradigm for Statistical Evidence* **University of Chicago**
- Schimmack, U. (2015): Replacing p-values with Bayes' Factors: A Miracle Cure for the Replicability Crisis in Psychological Science
- Scientific American (2015): Scientists Perturbed by Loss of Stat Tools to Sift Research Fudge from Fact
- Simonsohn, U., L. D. Nelson, J. D. Simmons (2014): p-Curve and Effect Size; Correcting for Publication Bias using only Significant Results *Perspectives on Psychological Science* **9 (6)** 666-681
- Stern, H. S. (2016): A Test by Any Other Name: p-values, Bayes' Factors, and Statistical Inference *Multivariate Behavioral Research* **51 (1)** 23-29
- Wasserstein, R. L., and N. A. Lazar (2016): The ASA's Statement on p-values: Context, Process, and Purpose *American Statistician* **70 (2)** 129-133
- Wetzels, R., D. Matzke, M. D. Lee, J. N. Rouder, G, J, Iverson, and E. J. Wagenmakers (2011): Statistical Evidence in Experimental Psychology: An Empirical Comparison using 855 t-Tests *Perspectives in Psychological Science* **6 (3)** 291-298
- Wikipedia (2019): p-value

Basel III Framework for Backtesting Exposure Models

Abstract

1. Standard Practices for IMM Back-testing: A central component of the Basel III (B3) standards is the *Sound Practices for Backtesting* (Basel Committee for Banking Supervision (2010)), i.e., a summary of the strict regulatory guidances on how to validate and back-test Internal Methods Models (IMM) for credit exposure.
2. Statistical Credit Exposure Backtesting Framework: In their work, Anfuso, Karyampas, and Nawroth (2017) define a comprehensive framework to backtest credit exposure models, highlighting the proposed features against the regulatory requirements.
3. Risk Factor Dynamical Evolution Backtesting: Their framework contains four main pillars. First is the *risk factor backtesting*, i.e., assessment of the forecasting ability of the stochastic differential equations (SDEs) used to describe the dynamics of the single factor.
4. Risk Factor Correlation Estimator Backtesting: Next is the *correlations backtesting*, i.e., the assessment of the statistical estimators used to describe the cross-asset evolution.
5. Representative Firm Portfolio Backtesting: Third is the *portfolio backtesting*, i.e., the assessment of the complete exposure model – SDEs + correlations + pricing – for portfolios that are representative of the firm's exposure.
6. Computation of the Capital Buffer: Last is the *computation of the capital buffer*, i.e., the extra amount of capital that the firm should hold if the model framework is not adequate – using the outcomes of the pillars above.
7. Distributional Tests for Collateralized/Uncollateralized: Anfuso, Karyampas, and Nawroth (2017) show with concrete examples in the cases of collateralized and uncollateralized models how to perform distributional tests with respect to different risk metrics.

8. Discriminatory Power Analysis across Forecasting Horizons: They produce discriminatory power analysis for all the tests introduced, providing exact methods to aggregate backtesting results across forecasting horizons.
9. Capital Remedies for Model Deficiencies: Most importantly, the third and the fourth pillars define a sound quantitative approach for computing capital remedies for potential model deficiencies.

Introduction

1. Validating and Backtesting IMM: The central pillar of the Basel III (B3) document is the *Sound Practices for Backtesting* (Basel Committee for Banking Supervision (2010)), i.e., a summary of strict regulatory guidelines on how to validate and backtest Internal Methods Models (IMM) for credit exposure.
2. European Basel III - The CRD4 Requirements: Similarly, a series of requirements for CRD4 – the European equivalent of B3 – indicate and define backtesting and validation as core component for good governance of IMM firms.
3. Importance of IMM for CVA: From a dealer perspective, the new regulatory changes introduced with B3 – e.g. CVA capital charge – have stressed even more the importance of IMM in making the capital costs of the businesses sustainable.
4. Importance of Backtesting for IMM: At the same time, the increasing complexity of the capital framework requires a thorough approach to the validation and the monitoring of the model performance.
5. Support from Regulators and Stake-holders: A sound backtesting methodology is therefore the key tool to both prove to the regulators the soundness of the models and to assure the stakeholders that the capital position of the firm is in sound modeling grounds.
6. Holistic Qualitative/Quantitative Model Assessment: The assessment of a model is a holistic process that has both qualitative and quantitative elements. While the former may have a

decisive weight for the choice of a given model many possible, the latter are the ones to be considered for backtesting.

7. Validation of the NULL Hypothesis: In particular, the performance of a model should be judged in terms of its forecasting ability. In statistical jargon, backtesting should address the question: *Can we reject the NULL hypothesis – i.e., the model – based on the available historical data*?
8. Statistical Testing of the Forecasting Ability: In the definition employed in this chapter, backtesting is therefore a set of statistical tests that measures the forecasting ability of the model using the data history available as comparison.
9. Assessment Metric - Aggregation of p-values: The final metric is based on the aggregation of the given p-values of the single tests, rejecting the model if an *a priori* determined threshold is breached.
10. Framework Features vs. Regulatory Guidelines: This chapter presents a complete framework to backtest credit exposure models. The next section gives a brief overview of the relevant metrics for the counterparty credit risk and summarizes the features of the framework against the new regulatory guidelines.
11. Complete Backtesting Cycle Details: In subsequent sections the methodology is presented in detail for the full backtesting cycle, i.e., the risk factor evolution models, the correlation models, and the portfolio exposure metrics. Later sections show how to compute the capital buffers based on the back-testing results. Finally, conclusions are drawn.

Basic Concepts and the Need for Backtesting

1. Credit Counterparty Exposure - Definition: Credit counterparty exposure is defined as the amount a dealer A could potentially loos in the case that its counterparty B defaults.

2. Estimating the CP Exposure Distribution: The exposure – from A's perspective – is computed from the forecasted distribution of prices of the financial contracts that constitute the portfolio of counterparty B at any future date.
3. Components required for the Exposure Estimation: The main building blocks required for the computation are the following: first, the scenario simulations for the underlying risk factors – generated with what is referred to here as the Risk Factor Evolution (RFE) models – and second, the pricing at each scenario to generate the $MTM(t)$ distribution at any future date t.
4. Expression for the Expected Positive Exposure: The relevant exposure metric from the regulatory perspective is the Expected (Positive) Exposure at time t which is defined as:

$$EE(t) = \mathbb{E}[MTM^{+}(t)]$$

where

$$MTM^{+}(t) = \max(MTM(t), 0)$$

5. RWA for CVA and Capital: The same exposure profile $EE(t)$ enters in the computation of the Risk Weighted Assets (RWA) of a given counterparty both for the CVA and for the default capital charges.
6. Modeling the Risk Factor Evolution: To compute the $MTM(t)$ distribution – and most importantly the $EE(t)$ profile – at any future time t one needs to forecast the evolution of the risk factor values. Those risk factors can also be dependent on each other – *correlation* assumption between the risk factors.
7. Loss Impact of RFE Mis-specification: The more accurately the RFE model is specified, the more realistic is the exposure calculation. If the RFE is mis-specified the exposure figure can be wrongly stated and the losses may occur with higher probability than expected in the case of counterparty defaults.
8. Risk Neutral Measure for the CVA: As observed in Kenyon and Stamm (2012), there is a potential for divergence in the choice for calibration for the RFE models. On the one hand,

the $EE(t)$ profile used for CVA calculation – a *price* – should be based on market calibration.

9. Historical Measure for Capital Charges: On the other hand, default charges require a forecast in the real-world measure and therefore a historical calibration would be most suitable.
10. Complication from the Dual Measures: The backtesting methodology described in the following is agnostic to the choice of the model calibration. Nevertheless, by construction, the regulatory requirements for backtesting models are addressed generally by historical models. In view of Anfuso, Karyampas, and Nawroth (2017), this apparent dichotomy is one of the key quantitative challenges for the industry after Basel III.
11. Regulatory Valuation Validation Framework: Following the guidances from regulators, Anfuso, Karyampas, and Nawroth (2017) define a framework that has four main pillars.
12. Risk Factor Dynamics Backtesting: The first is Risk Factor backtesting, i.e., the assessment of the forecasting ability of the Stochastic Differential Equations (SDEs) used to describe the dynamics of the single risk factors. It can be seen that the calibration of the SDE – market implied or historical – has a crucial influence on this assessment.
13. Risk Factor Correlation Backtesting: Next is the correlations backtesting, i.e., the assessment of the estimators used to model the cross-asset evolution.
14. The Representative Portfolio Backtesting: Third is the portfolio backtesting, i.e., the assessment of the complete exposure model – RFEs + Correlations + Pricing – for portfolios that are representative of the firm's exposure.
15. Model Reserve Capital Buffer Calculation: The computation of the capital buffer, i.e., the extra amount of capital that the form should hold if the model framework is not adequate – see the outcomes of the three pillars above – is the final pillar.
16. Diagnostic vs. Deficiency Remedy Pillars: The first three pillars are diagnostic whereas the fourth comes as a remedy for potential deficiencies in the exposure models.
17. Regulatory Guidance vs. Framework Response: The section below summarizes the relevant regulatory guidances for backtesting and how the framework in this chapter addresses them. The specifics of the proposed solutions are described in more detail in the sections below.

Regulatory Guidances

1. Guidance #1: The performance of the market risk factor should be validated using backtesting. The validation must be able to identify poor performances in individual risk factors.
    a. Methodology => The forecasting capability of the RFE models and their calibrations is back tested at multiple backtesting horizons, making use of different distributional tests.
    b. Compliance => Full.
2. Guidance #2: Validation of the EPE models and all the relevant models that input into the calculation of the EPE must be made using forecasts initialized on a number of historical dates.
    a. Methodology => The sampling of the backtesting is on a bi-weekly frequency, spanning all the available data history. Since the sampling is very close, no statistical bias due to the selection of the sampling frequency is introduced.
    b. Compliance => Full.
3. Guidance #3: Historical backtesting on representative counterparty portfolios and market risk factors must be part of the validation process. At regular intervals, as dictated by its supervisor, the dealer must conduct backtesting on a number of representative counterparty portfolios and the market risk factor models. The representative portfolios must be chosen based on the sensitivity to the material risk factors and correlations to which a dealer is exposed.
    a. Methodology => The portfolio backtesting is performed with suitable metrics that do not penalize the conservative estimates of the EPE. Representative counterparties can be chosen by a given dealer based on their RWA contributions.
    b. Compliance => Fully compliant from a methodology perspective. The dealer should additionally ensure that the selected counterparties are representative.

4. Guidance #4: Backtesting of the EPE and all the relevant models that input into the calculation of the EPE must be based on the recent performance.
    a. Methodology => The framework is agnostic to the data history that is selected, since that is an input. It can therefore be applied to assess recent and longer-term performances – though the statistical significance of the results will not be equivalent given the different amount of historical realizations that are being backtested.
    b. Compliance => Full.
5. Guidance #5: The frequency with which the parameters of an EPE model are updated needs to be assumed as part of the ongoing validation process.
    a. Methodology => The calibration is fully accounted for the RF, the correlations, and the portfolio backtesting.
    b. Compliance => Full.
6. Guidance #6: Dealers need to unambiguously define what constitutes acceptable and unacceptable performance for their EPE models and the models that input into the calculation of the EPE and have a written policy in place that describes how unacceptable performance will be remediated.
    a. Methodology => The framework gives a quantitative probabilistic interpretation of the performance that allows unambiguous acceptance or rejection of a model.
    b. Compliance => Fully compliant from a model perspective. The dealer should additionally ensure that the acceptance threshold is sufficiently conservative.
7. Guidance #7: IMM firms need to conduct hypothetical portfolio backtesting that is designed to test risk factor model assumptions, e.g., the relationships between the tenors of the same risk factors, and the modeled relationships between the risk factors.
    a. Methodology => The chapter does backtest the model correlation assumptions with a coherent extension of the methodology applied to the single risk factors.
    b. Compliance => Full.
8. Guidance #8: Firms must backtest their EPE models and all relevant models that input into the calculations of the EPE out to long horizons of at least one year.
    a. Methodology => Multiple horizons - shorter and longer than one year - are backtested at every level of granularity, i.e., RFs, correlations, and portfolio exposures.

   b. Compliance => Full.

9. Guidance #9: Firms must validate their EPE models and all relevant models that input into the calculation of the EPE out to the time horizons commensurate with the maturity of trades covered by the IMM waiver.
   a. Methodology => Multiple horizons commensurate with the maturity if the trades are backtested at every level of granularity, i.e., RF's, correlations, and portfolio exposures.
   b. Compliance => Full.
10. Guidance #10: Prior to the implementation of a new EPE model or a new model that inputs into the calculation of the EPE, a dealer must carry out a beck testing of the EPE model and all relevant models that input into the calculation of the EPE at a number of distinct time horizons using historical data for movements in the market risk factors for a range of historical periods covering a wide range of market conditions.
    a. Methodology => The framework described, because of its granularity and modularity, can be applied for periodic regulatory back testing as well as initial validation of a given model.
    b. Compliance => Full.

RF Backtesting: The Backtesting Construction for Collateralized and Uncollateralized Models

1. Basic Components of the Framework: The RFE models are the most atomic components of the exposure framework.
2. Cross Horizon RF Backtesting: As for regulatory guidances 1, 8, and 9, their performance should be assessed for different forecasting horizons and the predicted distributions should be consistent with the realized history of the corresponding risk factors.

3. Probability Integral Transform (PIT) Definition: The statistical tool that is the basis for the backtesting is the Probability Integral Transform (PIT – Gunther, Diebold, and Tay (1998), Kenyon and Stamm (2012)) defined as

$$F(r_n) = \int_{-\infty}^{r_n} \phi(x)dx$$

where r_n is the realization of the given random variable and $\phi(\cdot)$ is its predicted distribution.

4. Application to the RF Distribution: It is clear that when one applies PIT to a set of i.i.d. variables r_n using the correct distribution of r_n the transformed set

$$y_n = F(r_n)$$

is uniformly distributed.

5. Statistical Metric for Mismatch/Departure: The distance between the transformed set y_n and $U[0,1]$ distribution – in a statistical sense – can therefore be used as a goodness of the model $\phi(\cdot)$ to describe the random variable r_n.
6. RF Market Generation and Evolution: In practice, the input for RFE backtesting analysis is the time series of the given risk factor and the model – the SDE and its calibration – used to describe its evolution.
7. PIT Map onto $[0,1]$ Space: The PIT for the predicted distribution can be used to map the realized values of a risk factor – or their variations, see below – to a set of values in the interval $[0,1]$.
8. The Transformed Set Model Performance Assessment: The transformed set is then is used to assess the model performance by applying the standard distribution tests.
9. Collateralized vs. Uncollateralized Time Grids: For collateralized and uncollateralized models the construction differs because of the presence of the multiple time scales. In both cases a grid of sampling points t_k is used.

10. Criteria for Time Grid Selection: The sampling points define the origin of the backtesting experiment and they should be on a sufficiently fine grid and for a sufficiently long time series so as to ensure the following:
    a. An acceptable discriminatory power for the test – in relation to the size of the data history – and:
    b. An absence of significant statistical bias caused by the sampling sequence – if the sampling points are too distant, the backtesting results from an equivalent sequence with different arbitrary starting point may differ.
11. Horizon used for Backtesting: The backtesting is carried out for an arbitrary set of horizons $\{h_1, \cdots, h_n\}$ and for every h_i a single result is produced.
12. Horizon Matching Firms' Exposure Structure: The choice of the set $\{h_1, \cdots, h_n\}$ should be so as to reflect the portfolio structure of the firm.
13. Uncollateralized RF Model - Time Scale: In case of the uncollateralized RF models the horizon h_i is the only timescale. At

$$t = t_k$$

    the forecast forward distribution for

$$t = t_k + h_i$$

    is constructed based on he given RF model, and on the filtration $\mathcal{F}(t_k)$ at t_k.
14. PIT Based Transform to the $[0, 1]$ Scale: Using the PIT transform – where $\phi(\cdot)$ is given by the forecast conditional distribution – the realized value of RF at

$$t = t_k + h_i$$

    is mapped to a value in the $[0, 1]$ scale.
15. Collateralized RF Model - MPoR Impact: For collateralized models the presence of the margin period of risk ($MPoR$) should be additionally accounted for. The MPoR is the time

required for the dealer to liquidate the collateral a given counterparty posted to finance its exposure.

16. Collateralized RF Model - Primary Focus: Therefore, the primary focus of a collateralized model is to describe the variation of the RF over the MPoR at any future horizon.
17. Distribution of the RF inside the MPoR: In the backtesting exercise $\phi(\cdot)$ is the forecast RF variation distribution in the interval $[t = t_k + h_i, t = t_k + h_i + MPoR]$ conditional on $\mathcal{F}(t_k)$ and the realized value is the historical variation of RF in the same interval.
18. PIT Transformations on RF Realizations: The result of the PIT transformation on the sampling sequence is a set of values $\mathcal{F}(r_{t_k})$ in the interval $[0, 1]$.
19. Statistical Properties of the PIT Transformation: At this stage the standard statistical tests are applied to check for the different properties of the RF distribution.
20. Enhanced CDF Generalized Distribution Metric: In particular one can introduce a generalized distance metric as

$$d_w = \int_{\Gamma}^{\Gamma} [F_n(x) - F(x)]^2 w(x) dx$$

where

$$\Gamma = [0, 1]$$

is the domain of $F(x)$, $F_n(x)$ is the empirical cumulative distribution function CDF of the $F(r_{t_k})$ values,

$$F(x) = x$$

is the CDF of the $U[0, 1]$ distribution, and $w(x)$ is a weight function that can be chosen so as to emphasize a given quantile domain. In Kenyon and Stamm (2012) it is suggested to link $w(x)$ to the portfolio structure.

21. Individual RF Realizations as Pillars: Conversely, in this chapter the risk factors are backtested independently and the choice of $w(x)$ is unrelated to the portfolio composition. The portfolio backtesting is a separate pillar as described above.
22. Testing across different Weight Distribution: The literature provides many different statistical tests to check the match between the two distributions.
23. Cramer von Mises vs. Anderson Darling: This chapter considers the well-known Cramer-von Mises – CVM with

$$w(x) = 1$$

and Anderson-Darling (AD)

$$w(x) = \frac{1}{F(x)[1 - F(x)]}$$

tests where the first focuses more on the center of the distribution whereas the latter focuses more on the tails.

24. Distance Metric as a Single Value: It can be seen that d_w is a single *distance* value obtained from the realized cost at $\mathcal{F}(r_{t_k})$.
25. Monte Carlo based p-Value Estimate: The final outcome of the backtesting should be a p-Value. Therefore, the single realization d_w is assigned a p-Value based on the construction of the corresponding test statistic – distribution of the outcomes of d_w – using Monte Carlo.
26. Generation of the Test Statistic Distribution: The test statistic for a given history, sampling frequency, and horizon is produced by repeating on a large number of simulated paths the back-testing calculation/mapping described above.
27. Calibrated RF Model Path Generation: The paths are generated with the same model/calibration which is going to be used for calculating their corresponding d_w's.
28. Horizon Realized RF Value Distribution: Therefore, the test statistic is the distribution of the expected the correct model.

29. p-value from the d_w Quantiles: With the computed quantile of the realized d_w the p-value for the back-testing can be derived.
30. Handling of the Overlapping Forecasting Horizons: Notice that this derivation of the statistic allows for overlapping forecasting horizons since the auto-correlation among the $F\left(r_{t_k}\right)$ is correctly reflected in the construction.
31. Backtesting Longer Dated Horizons: This feature is particularly useful in backtesting longer horizons – see guidance #9, e.g., long-dated inflation/IR trades – for which the data history is comparatively short in most of the cases.
32. Sample Test - CHFUSD Exchange Rate: Anfuso, Karyampas, and Nawroth (2017) show uncollateralized and collateralized backtesting results for the CHFUSD exchange rate where they apply the CVM backtest to the last 15 years of history with bi-weekly sampling frequency for different forecasting horizons

$$h_i = \{1m, 3m\ 1y\}$$

33. RFE Process Underlying the FX: The RFE is a Geometric Random Brownian Motion (GBM) with drift

$$\mu = 0$$

and volatility calibrated using a rolling 1Y window.

34. Backtesting Collateralized/Uncollateralized Exposures: For the considered case, both the uncollateralized and the collateralized cases

$$MPoR = 2w$$

backtesting gives acceptable results at all horizons.

35. Discriminatory Power PLUS Horizon Aggregation: The next sections analyze two further aspects that were mentioned above – the discriminatory power of the statistical tests and the aggregation of the backtesting results across RF's and horizons.

Discriminatory Power RF Backtesting

1. Data Size Impact on Assessment: The assessment of the RF models depends crucially on the amount of data history.
2. Model Mis-specification Sensitivity to Size: In case of availability of large data sets, the backtesting can resolve very tiny model mis-specifications.
3. Less Data - Easier Backtesting: Conversely if the data is too few the model uncertainty will be larger – especially for longer horizons – and it will be comparatively easier to pass backtesting.
4. Quantitative Analysis of Data Sizes: To have a quantitative understanding of the above described effect, Anfuso, Karyampas, and Nawroth (2017) ran backtesting analysis on the CVM and the AD tests on synthetic data made from 1000 paths for 15 years generated by the same stochastic model – Geometric Brownian Motion with annualize drift and volatility of

$$\mu = 0$$

and

$$\sigma = 10\%$$

5. Impact of Drift/Volatility Mis-specifications: For every one of these paths, the p-value is determined for

$$h_i = \{1m, 3m, 1y\}$$

and for different mis-specifications of μ and σ.

6. Distance Metric as Sensitivity Metric: For a given h_i the average of the p-values across the paths is an intuitive measure of the sensitivity of the test for the given data history w.r.t. a mis-specified model calibration – for the correct model

$$\langle p \rangle = 0.5$$

7. Illustration of the above Analysis: Anfuso, Karyampas, and Nawroth (2017) illustrate the above analysis by highlighting the correctly specified model.
8. Backtesting Sensitivity to Shorter Horizons: As is evident from their tables, the backtesting results are more sensitive at the shorter horizons because of the larger number of independent observations taken into account:

$$h_i = 1m \rightarrow 180$$

vs.

$$h_i = 15y \rightarrow 15$$

independent observations.

9. AD vs. CVM Comparative Performance: It can also be noticed that AD slightly out-performs CVM in detecting model mis-specifications for the constructed example.
10. Discriminatory Power Analysis Regulatory Reporting: In the documentation that a dealer should provide to the regulators for the backtesting of the IMM, the discriminatory power analysis is a useful complementary information to assess the tolerance of the backtesting methodology.

The Aggregation of Backtesting Results

1. Lowest Granularity of Backtesting: The general indication from the regulators is to show the RF backtesting results at the most granular level, i.e., single RF's and horizons.
2. Asset Class and Horizon Aggregation: Nevertheless, the analysis should be complemented with aggregated results that assess more holistically the performance of the exposure models – e.g., by asset class and/or including several horizons.
3. *Boot-strappable* RF Probability Framework: The RF backtesting scheme presented above allows for further aggregations within the same probabilistic framework.
4. Aggregation over multiple RF Horizons: The scheme considered here is for a single RF over multiple horizons, i.e., the aim is to produce a single aggregated result where the importance of the different horizons is given an arbitrary weighting function $\theta(i)$ with

$$\sum_i \theta(i) = 1$$

and

$$\theta(i) > 0 \; \forall \; i$$

5. Incorporating the Dealer Portfolio Exposure: This case is of relevance in the common situation where the same RF model is used for products/portfolios of very different maturities and the assessment of mode performance should encompass many different time scales.
6. Standard Deviation of the Test Statistic: It can be seen that for a given sampling frequency that the standard deviation of the test statistic distribution scales linearly with the horizon. This is a direct consequence of the auto-correlation among back-testing experiments at different sampling points that grows linearly with the length of the horizon.
7. Standard Deviation for GBM AD/CVM: Anfuso, Karyampas, and Nawroth (2017) illustrate this behavior for both AD and CVM tests where the standard deviation has been determined numerically for the case of GBM for different forecasting horizons.

8. Horizon Normalization of the Standard Deviation: As a consequence of the linearity, the test normalized test distances

$$\bar{d}_{w,i} = \frac{d_{w,i}}{h_i}$$

with the normalization given by the forecasting horizon h_i - are measured in equivalent units.

9. Realized Historical Standard Deviation Metric: One can therefore define a single distance

$$d_{w,AGG} = \sum_i \theta(i)\bar{d}_{w,i}$$

with the desired weightings across the horizons.

10. Realized Standard Deviation Metric: Therefore, the realized historical value for $d_{w,AGG}$ is derived straightforwardly from $\bar{d}_{w,i}$ and $\theta(i)$.
11. Path-wise Test Statistic Distribution: The corresponding distribution of the test statistic can be computed by applying path by path the given test at different horizons and aggregating $d_{w,AGG}$ using the equation above.
12. Equi-Weighted Horizon Discriminatory Analysis: Anfuso, Karyampas, and Nawroth (2017) show the discriminatory power analysis – shown in the previous section – for the case of equally weighted aggregation across horizons

$$\theta(i) = \frac{1}{N_h}$$

where N_h is the number of horizons considered.

Correlations Backtesting

1. Basel III CRD Correlation Backtesting Estimator: One of the novelties introduced by Basel III and CRD4 for backtesting is the explicit requirement to check the correlation estimator – see Guidance #7.
2. Capturing Bear Market Correlated Moves: This is an important ingredient of the exposure framework, especially in periods of extreme bear markets when correlations rise significantly (Nawroth, Anfuso, and Akesson (2014)).
3. Enhancing PIT for Correlations Testing: While the previous sections were about the performance of single RF models, this section defines a method to measure how well a correlated set of SDE's for multiple RF's describe the RF's co-movements. As will be seen the PIT methodology can be suitably generalized for this purpose.
4. Backtesting Pair-wise Correlations: For a set of N RF's the aim is to backtest $\frac{N(N-1)}{2}$ correlations among all the – upper or lower – off-diagonal pairs.
5. Correlation Matrix for N GBM's: To explain the methodology consider the simplest model framework, i.e., N GBM's with a given – e.g. historical – calibration for $\vec{\mu}$, $\vec{\sigma}$, and the correlation matrix $[\rho]$.
6. Synthetic Consolidated Random Factor Consolidation: Defining the process for RF_i as

$$RF_i(t) = RF_i(0)e^{\mu_i t+\sigma_i W(t)-\frac{1}{2}{\sigma_i}^2 t}$$

for every non-equivalent pair $\{i,j\}$ one can introduce the synthetic RF Z_{ij} as follows:

$$Z_{ij}(t) = RF_i(t)^{\frac{1}{\sigma_i}}RF_j(t)^{\frac{1}{\sigma_j}}e^{\frac{1}{2}(\sigma_i+\sigma_j)t-\frac{1}{2}(2+2\rho_{ij})-\left(\frac{\mu_i}{\sigma_i}+\frac{\mu_j}{\sigma_j}\right)t}$$

7. Volatility of the Synthetic Random Factor: By construction Z_{ij} is a drift-less GBM with volatility given by

$$\sigma_{ij} = \sqrt{2 + 2\rho_{ij}}$$

8. Computing the Historical Z_{ij} Realizations: The historical realizations of Z_{ij} can be obtained at every sampling point using the estimators for the marginal distributions of RF_i and RF_j - i.e., the volatilities and the drifts $\sigma_i(t_k)$, $\sigma_j(t_k)$, $\mu_i(t_k)$, and $\mu_j(t_k)$, and the correlations among RF_i and RF_j.
9. Z_{ij} Volatility Measures ρ_{ij}: Z_{ij} can be therefore backtested as was done for single RF's, but its volatility is a direct measure of the correlation to be verified.
10. Inadequate Marginal Distribution Estimators: It can be seen that if the estimators of the marginal distributions are inadequate Z_{ij} is likely to fail backtesting independently of ρ_{ij}.
11. Consequence of Inadequate Marginal Estimations: This feature is not a drawback but rather a desirable property given the regulatory purpose of the backtesting analysis.
12. Need for Valid Marginal RF's: The RFE models are perceived as the atomic components of the exposure framework while the correlations are the second layer.
13. Valid Z_{ij} Estimators and Invalid Marginals: Whenever the underlying RF models fail the information on the correlation performance is of little value from a regulatory perspective.
14. SnP500 vs CHFUSD Correlation Tests: In their illustrations Anfuso, Karyampas, and Nawroth (2017) present uncollateralized backtesting results for SnP500 index and CHFUSD exchange rate.
15. GBM Martingale vs 1Y Rolling Volatility: They consider a correlated GBM with

$$\mu = 0$$

and the correlation ρ estimated using a 1Y rolling window.

16. Negative Correlation between SnP00 and USDCHF: The correlation is mostly negative, especially in correspondence of the lows of the SnP500.
17. Multi Horizon Correlations Backtesting: At the three-time horizons considered

$$h_i = \{1m, 3m, 1y\}$$

and for a data history of 15 years the correlation model passes backtesting at

$$CL = 99\%$$

- the two tickers pass RF backtesting independently for the same set of horizons. Results for the single factors are discussed earlier.

18. Reusing Single Factor Discriminatory Power Analysis: The mapping to the single RF problem allows the inheritance of all the results derived in that context – e.g., the discriminatory power analysis can be obtained as for single RF as is illustrated by Anfuso, Karyampas, and Nawroth (2017).
19. Collateralized/Uncollateralized Backtesting and Metric Aggregation: In particular the collateralized and the uncollateralized models can both be backtested based on the single RF factor scheme, and the correlation backtesting results can be aggregated at different horizons using the method discussed in the previous section.
20. Aggregation across a Single Correlator: Given the large number of entries in the correlation matrix for a given scenario, it is very convenient also to aggregate results across correlation elements – to a given forecasting horizon.
21. Block Level Correlation Metric Aggregation: A powerful visualization of the correlation testing is, e.g., aggregation by asset classes and the assignment of a given p-value for every block of the correlation matrix.
22. Block Level CVM/AD PIT: To obtain such a result, the aggregated distance for a given subset of the correlation matrix

$$\Omega = \{i \in \alpha, j \in \beta\}$$

can be obtained by applying the chosen distribution tests (CVM or AD) to the union of all the PIT's of the synthetic RF's

$$Z_{ij} \in \Omega$$

23. Model Comparison using the Departure Metric: The corresponding test statistic distribution can be derived with the model vs. model approach described in the previous sections, considering the correlated paths of the RF's in every scenario in Ω.
24. Cross RF Aggregation Technique: Anfuso, Karyampas, and Nawroth (2017) provide a detailed illustration of such an example of the aggregation methodology.

Portfolio Backtesting

1. Backtesting the Underlying RF's: The discussion so far has focused on the backtesting of the underlyings.
2. Primary Focus of the Regulators: However, the primary regulatory focus is on the performance on the overall regulatory framework, i.e., the ability of the dealer's IMM models to assess the RWA – i.e., the $EE(t)$ profile and hence the capital – accurately or conservatively – see Guidance #3.
3. RF vs. Regulatory Backtesting: It is obvious that the portfolio backtesting is conceptually different from the RF's and the correlations backtesting precisely owing to the above statement.
4. Tolerance for Overstating the Exposure: While in the case of individual RF's and correlations the model that is systematically understating or overstating a certain quantile domain is expected to fail, for portfolios and model feature that leads to systematically overstating the $EE(t)$ should not be penalized – at least from a regulatory perspective.
5. Asymmetry Adjusted PIT Capital Testing: The proposal of the third pillar of the framework – the portfolio backtesting – is again based on the PIT, with suitable modifications that account for the asymmetry discussed above.
6. Steps Involved in the Methodology: The methodology applies to both collateralized and uncollateralized books, and comprises of the following steps.

7. Scheme Based Counterparty Identification: The identification of a set of counterparties is based on, e.g., the B3 Default Capital RWA.
8. PIT for MM and r_t:The construction of the empirical uniform with the PIT using

$$F(r_n) = \int_{-\infty}^{r_n} \phi(x)dx$$

where $\phi(\cdot)$ is given by the forecasted MTM distribution for uncollateralized counterparties – or by the ΔMTM distribution for the collateralized ones – and r_t by the realized values for uncollateralized counterparties – or by realized MTM variation for collateralized ones.

9. Backtesting the Horizon-Appropriate Portfolio: At every sampling point

$$t = t_k$$

the composition of the portfolio should be the correct historical one. If no information on the historical trade composition is available one can simply backtest the current portfolio in a manner similar to what is discussed in the next section. The realized MTM value is obtained by re-pricing the same portfolio at

$$t = t_k + h$$

10. Statistically Conservative Test Portfolio #1: The statistical analysis of the $F(r_{t_k})$ set based on a test with a notion of conservatism embedded is referred to as the conservative portfolio CPT.
11. Statistically Conservative Test Portfolio #2: The CPT is defined as

$$CPT = \int_{\Gamma}^{\Gamma} [\max(F(x) - F(x_n), 0)]^2 w(x) dF(x)$$

where Γ and $F(x)$ are defined as for

$$d_w = \int_{\Gamma}^{\Gamma} [F_n(x) - F(x)]^2 w(x) dF(x)$$

where $F_n(x)$ is the empirical CDF from the forecasted MTM distribution and $w(x)$ is a weight function.

12. Affirmation of the Conservative Behavior: For a given quantile the $\max(\cdot, 0)$ function ensures that no test distance is accrued when the empirical uniform is *more conservative* than the theoretical one.
13. Considering Strictly Positive Exposures: In the exposure language if

$$F_n(x) < F(x)$$

the contribution of the quantile x to EE is greater or equal to the value from the exact model – equality holds in the case of x being negative when the contribution is zero.

14. Sample vs. Distribution Departure Characteristic: As shown in Anfuso, Karyampas, and Nawroth (2017), in such cases as the mis-specification of the volatility of the MTM distribution, $F(x)$ and $F_n(x)$ can have multiple crossing points.
15. Low/High MTM Volatility Impact: This implies that the lower quantiles of the estimation may be conservatively estimated while the higher values are below the correct values for low MTM volatilities. The inverse is true for high MTM volatilities.
16. Mean of the MTM Distribution: Additionally, the EE sensitivity of the MTM volatility depends on the mean of the MTM distribution.
17. IMT/ATM/OTM MTM Characteristic: Again, as shown in Anfuso, Karyampas, and Nawroth (2017), in the two limiting cases of the MTM distribution – deep in and deep out of the money, the EE is almost independent of the MTM volatility – as opposed to the more intuitive linear dependence for an MTM distribution centered at 0.
18. Differential Ranking of the EE Quantiles: For the determination of $w(x)$ the different quantiles should be ranked – for e.g., - their importance to the EE must be quantified.

19. Metric for the Quantile Distribution Importance: Their relative contribution$\alpha(q)$ can be defined as

$$\alpha(q) = \underset{\Delta q \to 0}{Limit} \max\left(\frac{1}{\Delta q}\int_{\Phi^{-1}(q)}^{\Phi^{-1}(q+\Delta q)} x\phi(x)dx\,, 0\right)\cdot\frac{1}{EE} = \frac{1}{EE}\max(\Phi^{-1}(q), 0)$$

where $\phi(x)$ and $\Phi^{-1}(x)$ are the probability density function and the inverse of the cumulative density function of a reference MTM distribution.

20. Application of the Mean Value Theorem: The identity in the above equation can be derived by applying the mean value theorem in the limit

$$\Delta q \to 0$$

21. Normal MTM Distribution - Relative Contribution: It can be seen that in the case of a normal MTM distribution, $\alpha(q)$ has the following closed-form solution:

$$\alpha(q, \mu, \sigma) = \frac{\max\left(\Phi_{\mathcal{N}(\mu,\sigma)}{}^{-1}(q), 0\right)}{EE_{\mathcal{N}(\mu,\sigma)}} = \frac{\max\left(\mu + \sqrt{2\sigma^2}EF^{-1}(2q-2), 0\right)}{\mu\Phi_{\mathcal{N}(0,1)}\left(\frac{\mu}{\sigma}\right) + \sigma\phi_{\mathcal{N}(0,1)}\left(\frac{\mu}{\sigma}\right)}$$

where $EF^{-1}(x)$ is the inverse of the error function, and μ and σ are the mean and the volatility of the Normal Distribution.

22. Same μ but varying σ: Anfuso, Karyampas, and Nawroth (2017) show $\alpha(q, \mu, \sigma)$ for a set of normal MTM distributions with equal volatility but different reasons.
23. Shape/Slope Dependence on μ: The shape and the slope of α vary dramatically with the level of MTM.
24. Par Contract α Volatility Dependence: It can be seen though that for

$$\mu = 0$$

the standard case of collateralized portfolio $\alpha(q, \mu, \sigma)$ is independent of the volatility, i.e.

$$\alpha(q, \mu = 0, \sigma) = \bar{\alpha}(q)$$

25. Elliptical Distribution μ Scale Invariance: This additional scale invariance property is quite robust and holds beyond the normality assumption, i.e., for all members of the elliptical family.
26. Inadequacy of a Single $w(x)$: For real portfolios the level of MTM varies across a wide spectrum of values and it is unlikely that a single choice of $w(x)$ can be optimal across all cases.
27. Special Case - Fully Collateralized Portfolios: In their analysis their focus has been on an MTM distribution centered at 0 – given also its special relevance for collateralized portfolios – and the case

$$w(x) = \bar{\alpha}(x)$$

is considered.

28. Application of the Appropriate $w(x)$: The dealer can fine-tune its own representation $w(x)$ by looking, for e.g., to the historically realized MTM's for the set of portfolios to be backtested.
29. Test Statistic appropriate for Portfolio Backtesting: A final remark is in order for the derivation of the test statistic for portfolio backtesting.
30. Multiple Times/MTM Sequence Simulation: The exact methodology – outlined for the RF's and correlations – would imply that every portfolio would have to compute its test statistic using two nested Monte Carlo simulations – one for the path of the underlyings, and one to generate the conditional MTM distribution at every sampling point and along every path.
31. Feasibility of the Dual Sequence Simulations: This approach can become computationally very expensive and requires a sophisticated parallel implementation on a cluster for large portfolios.

32. Options for avoiding Dual Simulations: If that is not feasible the following two alternative options are available.
33. Portfolios for Proxying Key Trades: For every representative portfolio select the most relevant trade and backtest only for the selected set – deriving the statistic exactly but for a much smaller portfolio.
34. Portfolio for Proxying Key RF's: Derive a representative statistic – used for all representative portfolios – based on an archetypal portfolio with trade composition and level of auto-correlation across sampling dates that are representative across of the set of portfolios to be backtested.

Capital Buffer Calculation

1. Basel III Compliant Backtesting Assessment: The previous sections have shown how to run a Basel III compliant backtesting and produce granular assessments for different IMM components.
2. Operational/Capital Remediation to Model Under-performance: Once that diagnosis is completed, in the case of unsatisfactory performance, the regulators expect the following steps from the dealer:
    a. A feedback loop so as to improve the models based on backtesting results
    b. An intermediate remediation action to account for potential shortages of capital to account for model deficiencies.
3. Capital Buffer Calculation – Requirement: This section discusses the fourth pillar of the framework, i.e., the calculation of the capital buffer (CB).
4. Capital Buffer Calculation - Edge Cases: The CB should be able to efficiently interpolate the two limits of a perfect forecasting model and its opposite, i.e., the case of a completely inadequate estimation of the regulatory capital.

5. Capital Buffer Edge Case Estimates: In the former situation the CB is equal to 0, while in the latter the sum of the estimated RWA and the buffer is bounded by regulatory capital established with standard rules – the conservative regulatory guidances that non-IMM dealers should follow for the calculation of the RWA.
6. Penalty for Model Mis-specification: Additionally, regulators expect the capital buffer to be punitive, i.e.,

$$RWA_{WM} \times (1 + CB_{WM}) \geq RWA_{RM}$$

   where WM and RM stand for *wrong* and *right* models and the CB has been defined as the multiplicative factor be applied to the IMM RWA.
7. Mandatory Penalty for Model Mis-specification: The role of the above inequality is to ensure that the dealer does not have any advantage in using the *wrong* models, i.e., that there is a capital incentive in adopting an adequate exposure framework.
8. Uniqueness of the Capital Buffer Estimation: The features of the CB stated above do not characterize it in a unique way.
9. Dependence on the Backtesting Performance: In the current construction the CB is linked to the performance of the portfolio backtesting for a selected set of N representative counterparties and is calculated historically by comparing the forecasted against the realized exposure profiles for these counterparties.
10. Interpretation as IMM Model Corrector: The final number can be interpreted as a correction factor to be applied to the IMM capital so as to account for the mis-specifications in the determination of the $EEPE$, i.e., the IM component of the RWA.
11. Importance of the CP Representative Portfolios: The capital buffer is entirely based on the representative set but is then applied to the whole portfolio. Therefore, the representative ness of the selected counterparties is a necessary feature to obtain a meaningful capital buffer.
12. Definition of the Model Error Metric: For a given counterparty c the following error metric is defined:

$$\Delta EE_c(t_1, t_2) = \max(MTM_c(t_2), 0) - EE_c(t_1, t_2)$$

where $\max(MTM_c(t_2), 0)$ is the realized exposure at t_2 and $EE_c(t_1, t_2)$ is the EE at t_2 as forecast at t_1.

13. 1Y Average of the Error Metric: The average of $\Delta EE_c(t_1, t_2)$ over one year is given by

$$\Delta EE_c(t_1) = \mathbb{E}_{t_2 \in [t_1, t_1+1Y]}[\Delta EE_c(t_1, t_2)]$$

and is the error over the whole horizon of the profile that is relevant for the $EEPE$ calculation.

14. The Capital Buffer Backtesting Metric Setup: As discussed above the capital buffer should be dependent on the portfolio backtesting performance.
15. Error Threshold Based Weight Function: This input enters into the calculation of the capital buffer via the following weighting function:

$$\mathcal{T}_c = \min\left(\frac{\max(p_c - p_l, 0)}{p_u - p_l}, 1\right)$$

where p_c is the p-value of the portfolio backtesting at

$$h = 1Y$$

- the most relevant horizon for the regulatory capital – for the counterparty c and p_l and p_u are the given lower and upper thresholds, i.e., 95% and 99%.

16. $\mathcal{T}_c$ Behavior over the p-Range: $\mathcal{T}_c$ is 0 below p_l and increases monotonically from 0 to 1 in the interval $[p_l, p_u]$.
17. Backtest Range: Satisfactory to Sufficient: p_u can be seen as the failure threshold for the portfolio backtesting while p_l is the lower bound of the p-values region where the backtesting performance may be considered unsatisfactory.

18. Fluctuation of Capital Buffer Estimates: In practice the capital buffer is calculated by IMM dealers on a quarterly basis and it is a desirable feature to not have large fluctuations in its size.
19. Implicit Smoothening inside the p-Range: Defining the relevant buffer region as an interval instead of as a binary threshold, the value of the buffer can smoothly account for potential deteriorations of the backtesting performance, rather than jumping suddenly to a much higher value if, e.g., a few top counterparties cross the failure threshold.
20. Time Horizon Capital Buffer: Having introduced $\mathcal{T}_c$ one can calculate the buffer for a given historical t_i as a p-value weighted across the N representative portfolios:

$$K(t_i) = \frac{\sum_{c=1}^{N} \mathcal{T}_c \Delta EE_c(t_i)}{\sum_{c=1}^{N} EEPE_c(t_i)}$$

where the sum of the forecasted $EEPE$'s in the denominator makes $K(t_i)$ a unit-less estimate for the relative error in the $EEPE$ framework.

21. Averaging the Capital Buffer Threshold: As a last step $K(t_i)$ can be averaged over the available history with either overlapping or non-overlapping samples.
22. Caveat - Include Only Positive Averages: The final result for the capital buffer is given by

$$CB = \max\left(\mathbb{E}_{t_i}[K(t_i)], 0\right)$$

where the $\max(\cdot, 0)$ function ensures that only the positive corrections apply, i.e., the negative corrections are not applied to the RWA.

23. Illustration of Backtesting with Capital Buffer: Anfuso, Karyampas, and Nawroth (2017) provide a complete series of portfolio backtesting and capital buffer calculations for synthetic MTM paths generated with Brownian motion for a portfolio of 20 independent counterparties. This results in several observations.
24. CPT vs. CVM or AD: First the discriminatory power of CPT is lower because less information about the MTM distribution is processed by the test in comparison with CVM or AD.

25. Finiteness of the Capital Buffer: The capital buffer is finite even in the case of correct model specification as a consequence of the finiteness of the history and of the $\max(\cdot, 0)$ function in

$$CB = \max\left(\mathbb{E}_{t_i}[K(t_i)], 0\right)$$

26. Full Cycle Capital Buffer Estimation: A full cycle calculation of the capital buffer shows that the inequality of

$$RWA_{WM} \times (1 + CB_{WM}) \geq RWA_{RM}$$

is fulfilled and therefore the buffer accounts for the missing capital due to the use of the wrong model.

27. Incorporating the Regulatory Upper Limit: While Anfuso, Karyampas, and Nawroth (2017) do not include the upper bound, i.e., the mandatory regulatory capital upper bound, this can be easily fixed by imposing

$$EEPE = \min(EEPE_{TOTAL}, EEPE_{REGULATORY})$$

Conclusion

1. Framework for Basel III Backtesting: This chapter includes a complete framework to backtest IMM according to the new Basel III guidelines.
2. Component of the Backtesting Scheme: The methodology includes a diagnosis of the models contributing to the CCR exposure – RF's, correlations, and portfolios – and a remedy for the potential model deficiencies impact on the regulatory capital.

3. Applications of the Backtesting Framework: It can be seen that the very same framework can be used as a template for:
    a. The development phase and the criteria that a model should meet prior to regulatory submission for the IMM waiver
    b. The internal one-off validation of a given model
    c. The periodic – e.g., quarterly – backtesting that the dealer should provide the regulators
4. Dealer's Internal Model Governance Scheme: A unified approach for validation and model backtesting is strongly endorsed by Basel III and CRD4 as it greatly simplifies the internal governance of the dealer.
5. Pre-requisite for IMM Model Approval: Additionally, in the new regulatory environment, the model developers should account for the backtesting requirements at the earliest stage, since strong backtesting performance is a key pre-requisite for IMM waiver approval.
6. Closed/MC/Rule Based Schemes: From a model perspective the methodology described above can be applied equally to Monte Carlo, historical, or rule-based CCR engines.
7. Enhancement to the Underlying Model Dynamics: The illustrations were based on GBM given their simplicity and relevance across the industry. Nevertheless, any other model can be backtested following the same logical steps.
8. Availability of Historical Data as a Shortcoming: In all cases the main bottleneck for a sound backtesting is the availability of sufficient historical data for statistically significant results.

References

- Anfuso, F., D. Karyampas, and A. Nawroth (2017): A Sound Basel III Compliant Framework for Backtesting Credit Exposure Models
- Basel Committee on Baking Supervision (2010): Sound Practices for Back-testing Counterparty Credit Risk Models

- Kenyon, C., and R. Stamm (2012): *Discounting, LIBOR, and Funding: Interest Rate and Credit Pricing* **Palgrave Macmillan**
- Nawroth, A., F. Anfuso, and F. Akesson (2014): Correlation Breakdown and the Influence of Correlations on VaR

Initial Margin Backtesting Framework

Abstract

1. Mandatory Margins for OTC Transactions: The introduction of mandatory margining for bilateral OTC transactions is significantly affecting the derivatives market, particularly in light of the additional funding costs that financial institutions could face.
2. Initial Margin Forecast Models Backtest: This chapter details the approach by Anfuso, Aziz, Loukopoulos, and Giltinan (2017) for a consistent framework, applicable equally to cleared and non-cleared portfolios, to develop and backtest forecasting models for initial margin.

Introduction

1. BCBS-IOSCO Mandatory Margining Guidelines: Since the publication of the new Basel Committee on Banking Supervision and the International Organizations of Securities Commissions (BCBS-IOSCO) guidance for mandatory margining for non-cleared OTC derivatives (Basel Committee on Banking Supervision (2015)) there has been a growing interest in the industry regarding the development of dynamic initial margin models (DIM) – see, for example, Green Kenyon (2015), Andersen, Pykhtin, and Sokol (2017b). By *DIM model* this chapter refers to any model that can be used to forecast portfolio initial margin requirements.

2. Protection Afforded by BCBS-IOSCO: The business case for such a development is at least two-fold. First, the BCBS-IOSCO IMR (B-IMR) rules are expected to protect against potential future exposure at a high-level of confidence (99%) and will substantially affect funding costs, XVA, and capital.
3. IM and VM Based Margining: Second, B-IMR has set a clear incentive for clearing; extensive margining in the form of variation margin (VM) and initial margin (IM) is the main element of the central counter-party (CCP) risk management as well.
4. IMR Impact on Bilateral + Cleared: Therefore, for both bilateral and cleared derivatives, current and future IMR significantly affects the probability and the risk profile of a given trade.
5. B-IMR Case Study - Performance Evaluation: This chapter considers B-IMR as a case study, and shows how to include a suitably parsimonious DIM model on the exposure calculation. It also proposes an end-to-end framework and also defines a methodology to backtest model performance.
6. Organization of this Chapter: This chapter is organized as follows. First, the DIM model for forecasting future IMR is presented. Then methodologies for two distinct levels of back-testing analysis are presented. Finally, conclusions are drawn.

How to Construct a DIM Model

1. Applications of the DIM Model: A DIM model can be used for various purposes. In the computation of the counter-party credit risk (CCR), capital exposure, or credit valuation adjustment (CVA), the DIM model should forecast, in a path-by-path basis, the amount of posted and received IM at any revaluation point.
2. Path Specific IMR Estimation: For this specific application, the key ability of the model is to associate a realistic IMR to any simulated market scenario based on a mapping that makes use of a set of characteristics of the path.

3. RFE Dependence on the DIM: The DIM model is *a priori* agnostic to the underlying risk factor evolution (RFE) models to generate the exposure paths (as shall be seen, dependencies may arise, if for example, the DIM is computed on the same paths that are generated for the exposure).
4. Cross-Probability Measure IMR Distribution: It is a different story if the goal is to predict the IMR distribution (IMRD) at future horizons, either in the real-world P or the market-implied Q measures.
5. IMRD Dependence on the RFE: In this context, the key feature of the model is to associate the right probability weight with a given IMR scenario; hence the forecast IMRD also becomes a measure of the accuracy if the IMRD models (which ultimately determine the likelihood of different market scenarios).
6. P vs. Q Measure IMRD: The distinction between the two cases will become clear later on, in the discussion of how to assess model performance.
7. ISDA SIMM BCBS IOSCO IM: The remainder of this chapter considers the BCBS-IOSCO IM as a case study. For the B-IMR, the current industry proposal is the International Swaps and Derivatives Association Standard Initial Margin Model (SIMM) – a static aggregation methodology to compute the IMR based on first-order delta-vega trade sensitivities (International Swaps and Derivatives Association (2016)).
8. Challenges with SIMM Monte Carlo: The exact replication of SIMM in a capital exposure or an XVA Monte Carlo framework requires in-simulation portfolio sensitivities to a large set of underlying risk factors, which is very challenging in most production implementations.
9. Andersen-Pykhtin-Sokol IM Proposal: Since the exposure simulation provides the portfolio mark-to-market (MTM) on the default (time t) and closeout (time $t + MPoR$, where $MPoR$ is the *margin period of risk*) grids, Andersen, Pykhtin, and Sokol (2017b) have proposed using this information to infer path-wise the size of any percentile of the local $\Delta MTM(t, t + MPoR, Path_i)$ distribution, based on a regression that uses the simulated portfolio $MTM(t)$ as a regression variable.
10. Andersen-Pykhtin-Sokol Proposal Assumptions: The

$$\Delta MTM(t, t + MPoR) = MTM(t + MPoR) - MTM(t)$$

distributed is constructed assuming that no cash flow takes place between the default and the closeout. For a critical review of this assumption, see Andersen, Pykhtin, and Sokol (2017a).

11. Enhancing the Andersen-Pykhtin-Sokol Model: This model can be further improved by adding more descriptive variables to the regression, e.g., values at the default time of the selected risk factors of the portfolio.
12. Optimization: Re-using Exposure Paths: For the DIM model, the following features are desirable. First the DIM should consume the same number of paths as the exposure simulation, to minimize the computational burden.
13. DIM Optimization – B-IMR SIMM Reconciliation: Second, the output of the DIM model should reconcile with the known IMR value for

$$t = 0$$

i.e.

$$IM(Path_i, 0) = IMR_{SIMM}(0)$$

for all i.

14. Key Aspects of IOSCO/SIMM: Before proceeding, this section notes some of the key aspects of the BCBS-IOSCO margining guidelines, and, consequently, of the ISDA SIMM Model (International Swaps and Derivatives Association (2016)).
15. Andersen-Pykhtin-Sokol Proposal Assumptions: First, the $MPoR$ for the IM calculation of a daily margined counter-party is $10\ BD$. This may differ from the capital exposure calculation, in which, for example

$$MPoR = 20\ BD$$

if the number of trades in the portfolio exceeds $5{,}000$.

16. No Netting across the Asset Classes: Second, the B-IMR in the Basel Committee on Banking Supervision (2015) prescribes calculating the IM by segregating trades from different asset classes. Tis feature is reflected in the SIMM model design.
17. SIMM Methodology Market Volatility Independence: Finally, the SIMM methodology consumes trade sensitivities as its only inputs and has a static calibration that is not sensitive to market volatility.
18. Regression on the ΔMTM Distribution: For the IM calculation, the starting point is similar to that of Andersen, Pykhtin, and Sokol (2017a), i.e.
    a. A regression methodology based on path's $MTM(t)$ is used to compute the moments of the local $\Delta MTM(t, t + MPoR, Path_i)$ distribution, and
    b. $\Delta MTM(t, t + MPoR, Path_i)$ is assumed to be a given probability distribution that can be fully characterized by its first two moments – the drift and the volatility. Additionally, since the drift is immaterial over the $MPoR$ horizon, it is not computed and set to 0.
19. Quadratic Regressor for Local Volatility: There are multiple regression schemes that can be used to determine the local volatility $\sigma(i,t)$ The present analysis follows the standard American Monte Carlo literature (Longstaff and Schwartz (2001)) and uses a least-squares method (LSM) with a polynomial basis:

$$\sigma^2(i,t) = \mathbb{E}[\Delta MTM^2(i,t) \mid MTM(i,t)] = \sum_{k=0}^{n} a_{\sigma k} MTM^k(i,t)$$

$$IM_{R/P,U}(i,t) = \Phi^{-1}\big(0.99/0.01, \mu = 0, \sigma = \sigma(i,t)\big)$$

where R/P indicates received and poste, respectively. In this implementation, the n in

$$\sigma^2(i,t) = \mathbb{E}[\Delta MTM^2(i,t) \mid MTM(i,t)] = \sum_{k=0}^{n} a_{\sigma k} MTM^k(i,t)$$

is set equal to 2, i.e., a polynomial regression of order 2 is used.

20. Calculating the Unnormalized IM Value: The unnormalized posted and received $IM_{R/P,U}(i,t)$ and calculated analytically in

$$\sigma^2(i,t) = \mathbb{E}[\Delta MTM^2(i,t) \mid MTM(i,t)] = \sum_{k=0}^{n} a_{\sigma k} MTM^k(i,t)$$

$$IM_{R/P,U}(i,t) = \Phi^{-1}\big(0.99/0.01, \mu = 0, \sigma = \sigma(i,t)\big)$$

by applying the inverse of the cumulative distribution $\Phi^{-1}(x,\mu,\sigma)$ to the appropriate quantiles; $\Phi(x,\mu,\sigma)$ being the probability distribution that models the local $\Delta MTM(t, t + MPoR, Path_i)$.

21. Note on the Distributional Assumptions: The precise choice of Φ does not play a crucial role, since the difference in the quantiles among the distribution can be compensated in calibration be applying the appropriate scaling factors (see the $\alpha_{R/P}(t)$ functions below). For simplicity, in the below Φ is assumed to be normal.
22. Comparative Performance of the LSM: It is observed that the LSM method performs well compared to the more sophisticated kernel methods such as Nadaraya-Watson, which is used in Andersen, Pykhtin, and Sokol (2017a), and it has the advantage of being parameter free and cheaper from a computational stand-point.
23. Applying $t = 0, MPoR$ and SIMM Reconcilers: The next step accounts for the

$$t = 0$$

reconciliation as well as the mismatch between SIMM and the exposure model calibrations – see the corresponding items above.

24. De-normalizing using IM Scaling Parameters: These issues can be tackled by scaling $IM_{R/P,U}(i,t)$ with suitable normalization functions $\alpha_{R/P}(t)$:

$$IM_{R/P}(i,t) = \alpha_{R/P}(t) \times IM_{R/P,U}(i,t)$$

$$\alpha_{R/P}(t) = \left[1 - h_{R/P}(t)\right] \times \sqrt{\frac{10\ BD}{MPoR}} \times \left[\alpha_{R/P,\infty} + \left(\alpha_{R/P,0} - \alpha_{R/P,\infty}\right)e^{-\beta_{R/P}(t)t}\right]$$

$$\alpha_{R/P,0} = \sqrt{\frac{MPoR}{10\ BD}} \times \frac{IMR_{R/P,SIMM}(t=0)}{q\left(0.99/0.01\ , \Delta MTM(0, MPoR)\right)}$$

25. Differential Calibration for Posted/Received IM: In

$$\alpha_{R/P}(t) = \left[1 - h_{R/P}(t)\right] \times \sqrt{\frac{10\ BD}{MPoR}} \times \left[\alpha_{R/P,\infty} + \left(\alpha_{R/P,0} - \alpha_{R/P,\infty}\right)e^{-\beta_{R/P}(t)t}\right]$$

$$\beta_{R/P}(t) > 0$$

and

$$h_{R/P}(t) < 1$$

with

$$h_{R/P}(t=0) = 0$$

are four functions to be calibrated – two for received and two for posted IM's. As will become clearer later in this chapter, the model calibration generally differs for received and posted DIM models.

26. Scaling IM using RFE MPoR: In

$$IM_{R/P}(i,t) = \alpha_{R/P}(t) \times IM_{R/P,U}(i,t)$$

$$\alpha_{R/P}(t) = \left[1 - h_{R/P}(t)\right] \times \sqrt{\frac{10\ BD}{MPoR}} \times \left[\alpha_{R/P,\infty} + \left(\alpha_{R/P,0} - \alpha_{R/P,\infty}\right)e^{-\beta_{R/P}(t)t}\right]$$

$MPoR$ indicates the $MPoR$ relevant for the Basel III exposure. The ratio of $MPoR$ to $10\ BD$ accounts for the VM vs. IM margin period, and it is taken as a square root because the underlying models are typically Brownian, at least for short horizons.

27. Components of the $\alpha_{R/P}(t)$ Term: In

$$\alpha_{R/P,0} = \sqrt{\frac{MPoR}{10\ BD}} \times \frac{IMR_{R/P,SIMM}(t=0)}{q\left(0.99/0.01\ , \Delta MTM(0, MPoR)\right)}$$

$IMR_{R/P,SIMM}(t=0)$ are the $IMR_{R/P}$ computed at

$$t = 0$$

using SIMM; $\Delta MTM(0, MPoR)$ is the distribution of the MTM variations over the first $MPoR$; and $q(x, y)$ is a function that gives quantile x for the distribution y.

28. $t = 0$ chosen to match SIMM: The values of the normalization functions $\alpha_{R/P}(t)$ at

$$t = 0$$

are chosen in order to reconcile $IM_{R/P}(i, t)$ with the starting SIMM IMR.

29. Mean-reverting Nature of the Volatility: The functional form of $\alpha_{R/P}(t)$ at

$$t > 0$$

is dictated by what is empirically observed, as is illustrated by Anfuso, Aziz, Loukopoulos, and Giltinan (2017); accurate RFE models, in both P and Q measures, have either a volatility

term structure or an underlying stochastic volatility process that accounts for the mean-reverting behavior to the normal market conditions generally observed from extremely low or high volatility.

30. Reconciliation with Static SIMM Methodology: Since the SIMM calibration is static (independence of market volatility for SIMM), the

$$t = 0$$

reconciliation factor is not independent of the market volatility, and thus not necessarily adequate for the long-term mean level.

31. Volatility Reducing Mean-reversion Speed: Hence, $\alpha_{R/P}(t)$ is an interpolant between the

$$t = 0$$

scaling driven by $\alpha_{R/P,0}$ and the long-erm scaling driven by $\alpha_{R/P,\infty}$, where the functions $\beta_{R/P}(t)$ are the mean-reverting speeds.

32. Estimating $\alpha_{R/P,\infty}$ from the Long-End: The values of $\alpha_{R/P,\infty}$ can be inferred by a historical analysis of a group of portfolios, or it can be *ad hoc* calibrated, e.g., by computing a different $\Delta MTM(0, MPoR)$ distribution in

$$\alpha_{R/P,0} = \sqrt{\frac{MPoR}{10\ BD}} \times \frac{IMR_{R/P,SIMM}(t=0)}{q\left(0.99/0.01\ , \Delta MTM(0, MPoR)\right)}$$

using the long-end of the risk-factor implied volatility curves and solving the equivalent scaling equations for $\alpha_{R/P,\infty}$.

33. Interpreting the Haircut $h_{R/P}(t)$ Term: As will be seen below, the interpretation of $h_{R/P}(t)$ can vary depending on the intended application of the model.

34. $h_{R/P}(t)$ for Capital/Risk Models: For capital and risk models, $h_{R/P}(t)$ are two capital and risk functions that can be used to reduce the number of back-testing exceptions (see below) and ensure that the DIM model is conservatively calibrated.
35. $h_{R/P}(t)$ for the XVA Models: For XVA pricing, $h_{R/P}(t)$ can be fine-tuned – together with $\beta_{R/P}(t)$ - to maximize the accuracy of the forecast based on historical performance.
36. Lack of Asset Class Netting: Note that owing to the *No netting across Asset Classes* clause, the $IM_{R/P,x}(i,t)$ can be computed on a stand-alone basis for every asset class x defined by SIMM (IR/FX, equity, qualified and non-qualified credit, commodity) without any additional exposure runs. The total $IM_{R/P}(i,t)$ is then given by the sum of the $IM_{R/P,x}(i,t)$ values.
37. Historical vs. Computed IM Calibrations: A comparison between the forecasts of the DIM model defined in

$$\sigma^2(i,t) = \mathbb{E}[\Delta MTM^2(i,t) \mid MTM(i,t)] = \sum_{k=0}^{n} a_{\sigma k} MTM^k(i,t)$$

$$IM_{R/P,U}(i,t) = \Phi^{-1}\left(0.99/0.01, \mu = 0, \sigma = \sigma(i,t)\right)$$

$$IM_{R/P}(i,t) = \alpha_{R/P}(t) \times IM_{R/P,U}(i,t)$$

$$\alpha_{R/P}(t) = \left[1 - h_{R/P}(t)\right] \times \sqrt{\frac{10\ BD}{MPoR}} \times \left[\alpha_{R/P,\infty} + \left(\alpha_{R/P,0} - \alpha_{R/P,\infty}\right) e^{-\beta_{R/P}(t)t}\right]$$

$$\alpha_{R/P,0} = \sqrt{\frac{MPoR}{10\ BD}} \times \frac{IMR_{R/P,SIMM}(t=0)}{q\left(0.99/0.01\,, \Delta MTM(0, MPoR)\right)}$$

and the historical IMR realizations computed with the SIMM methodology is shown in Anfuso, Aziz, Loukopoulos, and Giltinan (2017) where alternative scaling approaches are considered.

38. Criteria Utilized in the Comparison: A comparison is performed at different forecasting horizons using 7 years of historical data, monthly sampling, and averaging among a wide representation of single-trade portfolios for the posted and the received IM cases.
39. $\mathcal{L}_1$ Error Metric Choice: For a given portfolio/horizon, the chosen error metric is given by $\mathbb{E}_{t_k}\left[\frac{|F_{R/P}(t_k+h)-G_{R/P}(t_k+h)|}{G_{R/P}(t_k+h)}\right]$ where $\mathbb{E}_{t_k}[\cdot]$ indicates an average across historical sampling dates – the definitions of $F_{R/P}$ and $G_{R/P}$ are contained below. Here and throughout this chapter, t_k is used in place of t whenever the same quantity is computed at multiple sampling dates.
40. Comparison of the Tested Universe: The tested universe is made up of 102 single-trade portfolios. The products considered, always at-the-money and of different maturities, include cross-currency swaps, IR swaps, FX options, and FX forwards – approximately 75% of the population is made up of

$$\Delta = 1$$

trades.

41. Calibrated Estimates of the Parameters: As is made evident by Anfuso, Aziz, Loukopoulos, and Giltinan (2017), the proposed term structure of $\alpha_{R/P}(t)$ improves the accuracy of the forecast by a significant amount – they also provide the actual calibration used for their analysis.
42. Conservative Calibration of the Haircut Function: Below contains further discussions on the range of values that the haircut functions $\mathrm{h}_{R/P}(t)$ are expected to take for a conservative calibration of DIM to be used for regulatory exposure.
43. Comparison with CCP IMR: Finally, as an outlook, Anfuso, Aziz, Loukopoulos, and Giltinan (2017) show the error metrics for the case of CCP IMR where the Dim forecasts are compared against the Portfolio Approaches to Interest Rate Scenarios (Pairs: LCH.ClearNet) and historical value-at-risk (HVaR; Chicago Mercantile Exchange) realizations.
44. Prototype Replications of CCP Methodologies: The realizations are based on prototype replications of the market risk components of the CCP IM methodologies.

45. Universe Used for the CCP Tests: The forecasting capability of the model is tested separately for Pairs and HVaR IMR as well as for 22 single-trade portfolios (IRS trades of different maturities and currencies). The error at any given horizon is obtained by averaging among 22×2 cases.
46. Accuracy of the Proposed Scaling: Without fine tuning the calibration any further, the time-dependent scaling $\alpha_{R/P}(t)$ drives a major improvement in the accuracy of the forecasts with respect to the alternative approaches.

How to Backtest a DIM Model

1. Assessing Model for Different Applications: The discussion so far has focused on a DIM model for B-IMR without being too specific about how to assess the model performance for different applications, such as CVA and margin valuation adjustment (MVA) pricing, liquidity coverage ratio/net stable funding ratio (LCR/NSFR) monitoring (Basel Committee on Capital Supervision (2013)), and capital exposure.
2. Estimating the IMR Distribution Accurately: As mentioned above, depending upon which application one considers, it may or may not be important to have an accurate assessment of the distribution of *the simulated IM requirements* value (IMRD).
3. Backtesting to measure DIM Performance: This chapter introduces two distinct levels of backtesting that can measure the DIM model performance in two topical cases:
    a. DIM applications that do not depend directly on the IMRD (such as capital exposure and the CVA), and
    b. DIM applications that directly depend on the IMRD (such as MVA calculation and LCR/NSFR monitoring).

    The methodologies are presented below, with a focus on the *P*-measure applications.

Backtesting DIM Mapping Functions (for Capital Exposure and CVA)

1. Review of the Monte-Carlo Framework: In a Monte-Carlo simulation framework, the exposure is computed by determining the MTM values of a given portfolio on a large number of forward-looking risk-factor scenarios.
2. Adequacy of Forecasts across Scenarios: To ensure that a DIM model is sound, one should verify that the IM forecasts associated with the future simulation scenarios are adequate for a sensible variety of forecasting horizons as well as initial and terminal market conditions.
3. Setting up a Suitable Backtesting Framework: A suitable historical backtesting framework so as to statistically assess the performance of the model by comparing the DIM forecast with the realistic exact IMR, e.g., in the case of B-IMR calculated according to the SIMM methodology – for a representative sample of historical dates as well as market conditions and portfolios.
4. Generic IMR of a Portfolio: Let us first define generic IMR of a portfolio p as

$$IMR = g_{R/P}\left(t = t_\alpha, \Pi = \Pi\left(p(t_\alpha)\right), \vec{M}_g = \vec{M}_g(t_\alpha)\right)$$

   The terms are as follows.

5. Posted/Received IMR Computation Algorithm: The functions g_R and g_P represent the exact algorithm used to compute the IMR for the posted and the received IM's, respectively (e.g., such as SIMM for B-IMR, or in the case of the CCP's, IM methodologies such as Standard Portfolio Analysis of Risk (SPAN), Pairs, or HVaR).
6. Date of the IMR Valuation:

$$t = t_\alpha$$

   is the time at which the IMR portfolio p is determined.

7. Portfolio Trade Population at t_α: $\Pi\big(p(t_\alpha)\big)$ is the trade population of portfolio p at time t_α.
8. Market State Information at t_α: $\vec{M}_g(t_\alpha)$ is a generic state variable that characterizes all of the

$$T \leq t_\alpha$$

market information required for the computation of the IMR.

9. DIM Forecast of the Portfolio: Similarly, the DIM forecast for the future IMR of a portfolio p can be defined as

$$DIM = f_{R/P}\left(t_0 = t_k, t = t_k + h, \vec{r}, \qquad \Pi = \Pi\big(p(t_k)\big), \vec{M}_{DIM} = \vec{M}_{DIM}(t_k)\right)$$

The terms are as follows.

10. Posted/Received DIM Computation Algorithm: The functions f_R and f_P represent the DIM forecast for the posted and the received IM's, respectively.
11. Date of the DIM Forecast:

$$t_0 = t_k$$

is the date time at which the DIM forecast is computed.

12. Horizon of the DIM Forecast:

$$t = t_k + h$$

is the time for which the IMR is forecast – over a forecasting horizon

$$h = t - t_0$$

13. Predictor Set of Market Variables: $\vec{r}$ - the *predictor* – is a set of market variables whose forecasted values on a given scenario are consumed by the DIM models as input to infer the IMR.
14. $\vec{r}$ as Simulated Portfolio MTM: The exact choice of $\vec{r}$ depends on the DIM model. For the one considered previously, $\vec{r}$ is simply given by the simulated MTM of the portfolio.
15. Market State Information at t_k: $\vec{M}_{DIM}(t_k)$ is the generic state variable characterizing all the

$$T \leq t_k$$

market information required for the computation of the DIM forecast.

16. Portfolio Trade Population at t_k: $\Pi(\cdot)$ is defined as before.
17. Caveats around f_R and f_P: Despite being computed using the stochastic RFE models, f_R and f_P are not probability distributions, as they do not carry any information regarding the probability weight of a given received/posted IM value. $f_{R/P}$ are instead mapping functions between the set $\vec{r}$ chosen as predictor and the forecast value for IM.
18. Confidence Level Based DIM Calibration: In terms of $g_{R/P}$ and $f_{R/P}$ one can define exception counting tests. The underlying assumption is that the DIM model is calibrated at a given confidence level (CL); therefore, it can be tested as a $VaR(CL)$ model.
19. Model Conservatism Linked to CL: This comes naturally in the context of real-world P applications, such as capital exposure or liquidity monitoring, where a notion of model conservatism, and hence of exception, is applicable, since the model will be conservative whenever it understates (overstates) posted (received) IM.
20. The Portfolio Backtesting Algorithm Steps: For a portfolio p, a single forecasting day t_k, and a forecasting horizon h, one can proceed as follows.
21. t_k Estimate of the Forecast Functions: The forecast functions $f_{R/P}$ computed at time t_k are $f_{R/P}\left(t_0 = t_k, t = t_k + h, \vec{r},\ \Pi = \Pi\left(p(t_k)\right), \vec{M}_{DIM} = \vec{M}_{DIM}(t_k)\right)$ Note that $f_{R/P}$ depends exclusively on the predictor $\vec{r}$ –

$$\vec{r} = MTM$$

for the case considered above.

22. Impact of the Horizon on Predictor/Portfolio: The realized value of the predictor

$$\vec{r} = \vec{R}$$

is determined. For the model considered above, $\vec{R}$ is given by the portfolio value $p(t_k + h)$ where the trade population $\Pi\big(p(t_k + h)\big)$ at $t_k + h$ differs from t_k only because of portfolio aging. Aside from aging, no other portfolio adjustments are made.

23. Forecast Received/Posted IMR Estimate: The forecast values for the received and the posted IM's are computed as

$$F_{R/P}(t_k + h) = f_{R/P}\left(t_0 = t_k, t = t_k + h, \vec{r}, \quad \Pi = \Pi\big(p(t_k)\big), \vec{M}_{DIM} = \vec{M}_{DIM}(t_k)\right)$$

24. Forecast of the Received/Posted IM Estimate: The realized values for the received and the posted IM's are computed as

$$G_{R/P}(t_k + h) = g_{R/P}\left(t = t_k + h, \Pi = \Pi\big(p(t_k + h)\big), \vec{M}_g = \vec{M}_g(t_k + h)\right)$$

25. Exception Case: F/G Mismatch Conservatism: The forecast and the realized values are then compared. The received and the posted DIM models are considered independently, and a backtesting exception occurs whenever $F_R(F_P)$ is larger (smaller) than $G_R(G_P)$. As discussed above, this definition of exception follows from the applicability of a notion of model conservatism.
26. Detecting the Backtesting Exception History: Applying the above steps to multiple sampling points t_k one can detect back-testing exceptions for the considered history.
27. Dimensionality Reduction for the Comparison: The key step is the estimate of the posted/received IMR forecast, where the dimensionality of the forecast is reduced – from a

function to a value – making use of the realized value of the predictor, and, hence, allowing for a comparison with the realized IMR.

28. Determining the Test p-value using TVS: The determination of the test p-value requires additional knowledge of the Test Value Statistics (TVS), which can be derived numerically if the forecasting horizons are overlapping (Anfuso, Karyampas, and Nawroth (2017)).
29. Caveats behind Blind TVS Usage: In the latter situation, it can happen that a single change from one volatility regime to another may trigger multiple correlated exceptions; hence the TVS should adjust the beck-testing assessments for the presence of false positives.
30. Accuracy of the $\alpha_{R/P}(t)$ Scaling: The single trade portfolios seen earlier have been tested by Anfuso, Aziz, Loukopoulos, and Giltinan (2017) using the SIMM DIM models with the three choices of scaling discussed earlier. The results confirm the greater accuracy of the term structure scaling of $\alpha_{R/P}(t)$.
31. Accuracy in the Presence of Haircut: In fact, for the same level of the haircut function

$$h_{R/P}(t > 0) = \pm 0.25$$

positive/negative for posted/received – a much lower number of exceptions is detected.

32. Realistic Values for the Haircut: Anfuso, Aziz, Loukopoulos, and Giltinan (2017) also observe that, in this regard, for realistic diversified portfolios and calibration targets of

$$CL = 95\%$$

the functions $h_{R/P}(t)$ take values typically in the range of $10 - 40\%$.

33. Assumptions Underlying the Haircut Assumption: The range of values for $h_{R/P}(t)$ has been calibrated using

$$\beta_{R/P}(t) = 1$$

and

$$\alpha_{R/P,\infty}(t) = 1$$

Both assumptions are broadly consistent with historical data.

34. IOSCO results in Over-collateralization: Note also that the goal of the BCBS-IOSCO regulations is to ensure that the netting sets are largely over-collateralized as a consequence of:
    a. The high confidence level at which the IM is computed, and
    b. The separate requirements for IM and VM.
35. Impact of Over-collateralization: Hence, the exposure generating scenarios are tail events, and the effect on capital exposure of a conservative haircut applied to the received IM is rather limited in absolute terms.
36. Over-collateralization Impact on Exposure: This issue is demonstrated by Anfuso, Aziz, Loukopoulos, and Giltinan (2017) where the expected exposure (EE) at a given horizon t is shown as a function of $h_R(t)$ – the haircut to be applied to the received IM collateral – for different distributional assumptions on $\Delta MTM(t, t + MPoR)$.
37. Distribution Dependence on Haircut Functions: In particular, they compute the expected exposure for

$$h_R(t) = 0$$

and

$$h_R(t) = 1$$

indicating full IM collateral benefit or no benefit at all – and take the unscaled IM as the 99^{th} percentile of the corresponding distribution. For different classes of the ΔMTM distribution, the exposure reduction is practically unaffected up to haircuts of $\approx 50\%$.

Backtesting the IMRD for MVA and LCR/NSFR

1. MC Based DIM IMR Distributions: The same Monte Carlo framework can be used in combination with a DIM model to forecast the IMD at any future horizon – implicit here are the models in which the DIM is not always constant across the scenarios. The applications of the IMRD are multiple.
2. Some Applications using the IMRD: The following are two examples that apply equally to the cases of B-IMR and CCP IMR:
    a. Future IM funding costs in the P measure, i.e., the MVA
    b. Future IM funding costs in the Q measure, e.g., in relation to LCR and NSFR regulations (Basel Committee on Banking Supervisions (2013))
3. Numerically Forecasting the IMR Distributions: The focus here is on the forecasts on the P-measure – tackling the case of the Q-measure may require a suitable generalization of Jackson (2013). The main difference with the backtesting approach discussed above is that the new model forecasts are the numerical distributions of the simulated IMR values.
4. Scenario-specific IM Forecasting: These can be obtained for a given horizon by associating every simulated scenario with its corresponding IMR forecast, computed according to the given DIM model.
5. Posted/Received IMR Density CDF: Using the notation introduced previously, the numerical representations of the received/posted IMRD cumulative density functions (CDF's) of a portfolio p for a forecasting day t_k and a horizon h are given by

$$CDF_{R/P}(x, t_k, h) = \frac{\#\{v \in \mathbb{V} \mid v \leq x\}}{N_{\mathbb{V}}} \; \forall \, \vec{r}_\omega \in \Omega$$

$$\mathbb{V} = \left\{ f_{R/P}\left(t_0 = t_k, t = t_k + h, \vec{r}, \qquad \Pi = \Pi\big(p(t_k)\big), \vec{M}_{DIM} = \vec{M}_{DIM}(t_k)\right)\right\}$$

6. Terms of the CDF Expression: In

$$CDF_{R/P}(x, t_k, h) = \frac{\#\{v \in \mathbb{V} \mid v \leq x\}}{N_{\mathbb{V}}}$$

$N_{\mathbb{V}}$ is the total number of scenarios. In

$$\mathbb{V} = \left\{ f_{R/P}\left(t_0 = t_k, t = t_k + h, \vec{r}, \quad \Pi = \Pi\left(p(t_k)\right), \vec{M}_{DIM} = \vec{M}_{DIM}(t_k)\right) \forall\, \vec{r}_\omega \in \Omega \right\}$$

$f_{R/P}$ are the functions computed using the DIM model, $\vec{r}_\omega$ are the scenarios for the predictor – the portfolio MTM values in the case originally discussed, and Ω is the ensemble of $\vec{r}_\omega$ spanned by the Monte Carlo simulation.

7. Suitability of IMRD for Backtesting: The IMRD in this form is directly suited for historical backtesting using the Probability Integral Transformation (PIT) framework (Diebold, Gunther, and Tay (1998)).
8. Forecasting Horizon PIT Time Series: Referring to the formalism described in one can derive the PIT time series $\tau_{R/P}$ for a portfolio p for a given forecasting horizon h and backtesting history $\mathcal{H}_{BT}$ as:

$$\tau_{R/P} = CDF\left(g_{R/P}\left(t = t_k + h, \Pi = \Pi\left(p(t_k + h)\right), \vec{M}_g = \vec{M}_g(t_k + h)\right), t_k, h\right) \forall\, t_k \in \mathcal{H}_{BT}$$

9. Samples from the Actual IMR Algorithm: In the expression for $\tau_{R/P}$ above, $g_{R/P}$ is the exact IMR algorithm for the IMR methodology that is to be forecast – defined as

$$IMR = g_{R/P}\left(t = t_\alpha, \Pi = \Pi\left(p(t_\alpha)\right), \vec{M}_g = \vec{M}_g(t_\alpha)\right)$$

and t_α are the sampling points in $\mathcal{H}_{BT}$.

10. Probability of t_k-realized IMR: Every element of the PIT time series $\tau_{R/P}$ corresponds to the probability of the realized IMR at time $t_k + h$ according to the DIM forecast built at t_k.
11. Backtesting of the Portfolio Models - Variations: As discussed extensively in Anfuso, Karyampas, and Nawroth (2017) one can backtest $\tau_{R/P}$ using uniformity tests. In particular,

analogous to what was shown in Anfuso, Karyampas, and Nawroth (2017) for portfolio backtesting in the context of capital exposure models, one can use test metrics that do not penalize conservative modeling – i.e., models overstating/understating posted/received IM. In all cases the appropriate TVS can be derived using numerical Monte Carlo simulations.

12. Factors affecting the Backtesting: In this setup the performance of a DIM is not done in isolation. The backtesting results will be mostly affected by the following.
13. Impact of $\vec{r}$ on Backtesting: As discussed earlier, $\vec{r}$ is the predictor used to associate an IMR with a given scenario/valuation time point. If $\vec{r}$ is a poor indicator for the IMR, the DIM forecast will consequently be poor.
14. Mapping of $\vec{r}$ to IMR: If the mapping model is not accurate, then the IMR associated with a given scenario will be inaccurate. For example, the models defined in

$$\sigma^2(i,t) = \mathbb{E}[\Delta MTM^2(i,t) \mid MTM(i,t)] = \sum_{k=0}^{n} a_{\sigma k} MTM^k(i,t)$$

$$IM_{R/P,U}(i,t) = \Phi^{-1}\big(0.99/0.01, \mu = 0, \sigma = \sigma(i,t)\big)$$

$$IM_{R/P}(i,t) = \alpha_{R/P}(t) \times IM_{R/P,U}(i,t)$$

$$\alpha_{R/P}(t) = \big[1 - h_{R/P}(t)\big] \times \sqrt{\frac{10\ BD}{MPoR}} \times \big[\alpha_{R/P,\infty} + \big(\alpha_{R/P,0} - \alpha_{R/P,\infty}\big)e^{-\beta_{R/P}(t)t}\big]$$

$$\alpha_{R/P,0} = \sqrt{\frac{MPoR}{10\ BD}} \times \frac{IMR_{R/P,SIMM}(t=0)}{q\big(0.99/0.01\,, \Delta MTM(0, MPoR)\big)}$$

include scaling functions to calibrate the calculated DIM to the observed

$$t = 0$$

IMR. The performance of the model is therefore dependent on the robustness of this calibration at future points in time.

15. RFE Models used for $\vec{r}$: The models ultimately determine the probability of a given IMR scenario. It may so happen that the mapping functions $f_{R/P}$ are accurate but the probabilities for the underlying scenarios for $\vec{r}$ are misstated, and, hence, cause backtesting failures.
16. Differential Impact of Backtesting Criterion: Note that
    a. The choice of $\vec{r}$, and
    b. The mapping

$$\vec{r} \rightarrow IMR$$

are also relevant to the backtesting methodology discussed earlier in this chapter. RFE models used for $\vec{r}$, however, are particular to this backtesting variance, since it concerns the probability weights of the IMRD.

Conclusion

1. Framework to Develop/Backtest DIM: This chapter has presented a complete framework to backtest and develop DIM models. The focus has been on B-IMR and SIMM, and the chapter has shown how to obtain forward-looking IM's from the simulated exposure paths using simple aggregation methods.
2. Applicability of the Proposed Model: The proposed model is suitable for both XVA pricing and capital exposure calculations; the haircut functions in

$$\alpha_{R/P}(t) = \left[1 - h_{R/P}(t)\right] \times \sqrt{\frac{10\ BD}{MPoR}} \times \left[\alpha_{R/P,\infty} + \left(\alpha_{R/P,0} - \alpha_{R/P,\infty}\right)e^{-\beta_{R/P}(t)t}\right]$$

can be used to either improve the accuracy (pricing) or to ensure the conservatism of the forecast (capital).

3. CCR Capital using DIM Models: If a financial institution were to compute CCR exposure using internal model methods (IMM), the employment of a DIM could reduce the CCR capital significantly, even after the application of a conservative haircut.
4. Over-collateralization inherent in Basel SA-CCR: This should be compared with the regulatory alternative SA-CCR, where the benefits from over-collateralization are largely curbed (Anfuso and Karyampas (2015)).
5. Backtesting Methodology to Estimate Performance: As part of the proposed framework, this chapter introduced a backtesting methodology that is able to measure model performance for different applications of DIM.
6. Agnosticity of DIM to the Underlying IMR: The DIM model and the backtesting methodology presented are agnostic to the underlying IMR algorithm, and they can be applied in other contexts such as CCP IM methodologies.

References

- Andersen, L., M. Pykhtin, and A. Sokol (2017a): Re-thinking Margin Period of Risk **eSSRN**.
- Andersen, L., M. Pykhtin, and A. Sokol (2017b): Credit Exposure in the Presence of Initial Margin **eSSRN**.
- Anfuso, C., D. Aziz, K. Loukopoulos, and P. Giltinan (2017): A Sound Modeling and Backtesting Framework for Forecasting Initial Margin **eSSRN**.
- Anfuso, C., and D. Karyampas (2015): Capital Flaws *Risk* **27 (7)** 44-47

- Anfuso, C., D. Karyampas, and A. Nawroth (2017): A Sound Basel III Compliant Framework for Backtesting Credit Exposure Models **eSSRN**.
- Basel Committee on Banking Supervision (2013): Basel III: The Liquidity Coverage Ratio and Liquidity Risk Monitoring Tools
- Basel Committee on Banking Supervision (2015): Margin Requirements for Non-centrally Cleared Derivatives
- Diebold, F. X., T. A. Gunther, and A. S. Tay (1998): Evaluating Density Forecasts with Applications to Financial Risk Management *International Economic Review* **39 (4)** 863-883
- Green, A. D., and C. Kenyon (2015): MVA by Replication and Regression **arXiV**
- International Swaps and Derivatives Association (2016): ISDA SIMM Methodology
- Jackson, L. (2013): Hedge Backtesting for Model Validation *Risk* **25 (9)** 64-67
- Longstaff, F., and E. Schwartz (2001): Valuing American Options by Simulation: A Simple Least-Squares Approach *Review of Financial Studies* **14 (1)** 113-147

Model Risk Management Framework

Introduction

1. Widespread Use of Quantitative Methods: Banks rely heavily on models and quantitative analysis in most aspects of financial decision making (Federal Reserve (2011)).
2. Governing Agencies of the Banks: Unless otherwise indicated, *banks* refer to national banks and all other institutions for which the Office of Comptroller of the Currency is the primary supervisor, and to Bank Holding Companies, state member banks, and all other institutions for which the Federal Reserve Board is the primary supervisor.
3. Activities that Employ Quantitative Techniques: Banks routinely use models for a broad rang of activities, including under-writing credits; valuing exposures, positions, and instruments; managing risk; managing and safeguarding client assets; determining capital and reserve adequacy; and many other activities.
4. Expanding Use of Quantitative Schemes: In recent years, banks have applied models to more complex products, and with more complex scope, such as enterprise-wide risk management, while the markets in which they are used have also broadened and changed.
5. Regulatory Demands Fulfilled by Quantitative Means: Changes in regulation have spurred some of the recent developments, particularly the US regulatory capital rules for market, credit, and operational risk based on the framework developed by the Basel Committee for Banking Supervision.
6. Non-Regulatory Drivers of Quantitative Usage: Even apart from these regulatory considerations, however, banks have been increasing the use of data-driven, quantitative decision-making tools for a number of years.

7. Benefits and Costs of Models: The expanding use of models in all aspects of banking reflects the extent to which models can improve business decisions, but models also come with costs.
8. Costs of Model Development and Implementation: There is the direct cost of devoting resources to develop and implement models properly.
9. Indirect Costs - Underlying Models Usage: There is also the potential indirect cost of relying on the models, such as possible adverse consequences – including financial loss – of decisions based on models that are incorrect and misused.
10. Role of Model Risk Management: These consequences should be addressed by active management of model risk.
11. Implementing Effective Model Risk Management: This chapter describes the key aspects of effective model risk management.
12. Purpose and Scope of this Chapter: The next section explains the purpose and the scope of the guidance, following which an overview of the model risk management is provided.
13. Model Development, Implementation, and Use: The next section discusses robust model development, implementation, and use.
14. Effective Validation Framework Components: The section after that describes the components of an effective validation framework.
15. Policies and Controls for Governance: The penultimate section explains the salient features of good governance, policies, and controls, from the viewpoint of model development, implementation, use, and validation.
16. Concluding Remarks on the Guidance: The final section concludes.

Purpose and Scope

1. Guidance on Model Risk Management: The purpose of this chapter is to provide comprehensive guidance on model risk management.

2. Rigorous Model Validation and Backtesting: Rigorous model validation plays a critical role in model risk management; however, sound development, implementation, and use of models are also vital elements.
3. Model Risk Management Governance/Control: Furthermore, model risk management comprises governance and control mechanisms such as board and senior management oversight, policies and procedures, controls and compliance, and an appropriate incentive and organizational structure.
4. Importance of Model Validation: Previous guidance and other publications issued by the OCC and the Federal Reserve on the use od models pay particular attention to model validation.
5. Enhancements over Previous Supervisory Guidelines: Based on supervisory and industry experience over the past several years, this chapter expands on existing guidance – most importantly by broadening the scope to include all aspects of model risk management.
6. OCC Governance on Model Risk: For instance, OCC provided guidance on model risk, focusing on model validation, in OCC-2016 (30 May 2000), other bulletins, and certain subject matter booklets of the *Comptroller's Handbook.*
7. Federal Reserve's Guidance on Model Risk: The Federal Reserve issued SR Letter 09-01, *Application of Market Risk Rules in Bank Holding Companies and State Member Banks*, which highlights various concepts pertinent to model risk management, including standards for validation and review, model validation documentation, and backtesting.
8. Federal Reserve's Validation and MRM Guidance: The Federal Reserve's *Trading and Capital Markets' Activities Manual* also discusses validation and model risk management.
9. Validation Requirements for Subject Banks: In addition, the advanced approaches risk-based capital rules (12 CFR 3, Appendix C; 12 CFR 208, Appendix F; and 12 CRF 225, Appendix G) contain explicit validation requirements for subject banking organizations.
10. Supervisory Requirements over Model Risk: Many banks may already have in place a large portion of these practices, but all banks should ensure that their internal policies and procedures are consistent with the management principles and supervisory expectations contained in this guidance.

11. Practical Application of the Guidance: Details may vary from bank to bank, as the practical implementation of this guidance should be customized to be commensurate with the bank's exposures, its business activities, and the complexity and the model of its model use.
12. Variability in Applying the Guidelines: For example, steps taken in applying this guidance at a community bank using relatively few models of only moderate complexity might be significantly less involved than those at a larger bank where the use of models is more extensive or complex.

Overview of Model Risk Management

1. Complete Definition of a Model: For purposes of this chapter, the term *model* refers to a quantitative method, system, or approach that uses statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates.
2. Component Entities inside a Model: A *model* consists of three components; an information input component, which delivers the assumptions and data into the model; a processing component, which transforms inputs into estimates; and a reporting component, which translates the estimates into useful business information.
3. Diverse Applications of a Model: Models meeting this definition might be used for analyzing business strategies, informing business decisions, identifying and measuring risks, valuing exposures, instruments, or positions, conducting stress testing, assessing adequacy of capital, managing client assets, measuring compliance with internal limits, maintaining the formal control apparatus of the bank, meeting financial or regulatory reporting requirements, or issuing public disclosures.
4. Qualitative Inputs into a Model: The definition of a *model* also covers quantitative approaches whose inputs are partially or wholly qualitative or based on expert judgement, provided that the approach is quantitative in nature.

5. Control Process for Quantitative Approaches: While outside the scope of this guidance, more quantitative approaches used by banking organizations, i.e., those *not* defined as models according to this guidance, should also be subject to rigorous control process.
6. Models as Representative of Real-World Relationships: Models are simplified representations of real-world relationships among observed characteristics, values, and events.
7. Simplification - Both Incidental and Intentional: Simplification is inevitable due to the inherent complexity of those relationships, but also intentional, to focus attention on particular aspects considered to be the most important for a given model application.
8. Measuring the Qualities of a Model: Model quality can be assessed in many ways; precision, accuracy, discriminatory power, robustness, stability, reliability, to name a few.
9. Improving the Quality of Models: Models are never perfect, and the appropriate metrics of quality, and the effort that should be put into improving quality, depend on the situation.
10. Example - Precision/Accuracy vs. Discriminant: For example, precision and accuracy are important for models that forecast future values, while discriminatory power applies to models that rank order the risks.
11. Understanding Model's Capabilities and Limitations: In all situations, it is important to understand the model's capabilities and limitations, given its simplifications and assumptions.
12. The Danger of Model Risk: The use of models invariably presents model risk which is the potential for adverse consequences based on incorrect or misused model outputs and reports.
13. Some Consequences of Model Errors: Model risk can lead to financial loss, poor business or strategic decision making, or damage to a bank's reputation. Model risk occurs primarily for two reasons.
14. Model Errors at the Design Stage: The model may have fundamental errors and may produce inaccurate outputs when viewed against the design objective and the intended business use.
15. Elements in a Model Pipeline: The mathematical calculation and the quantification exercise underlying any model generally involves application of the theory, choice of sample design and numerical routines, selection of inputs and estimation, and implementation in information systems.

16. Locations of Model Errors: Errors can occur at any stage from model design through implementation.
17. Shortcuts, Simplifications, and Model Approximations: In addition, shortcuts, simplifications, and approximations used to manage complicated problems could compromise the integrity and the reliability of the outputs from those calculations.
18. Errors in Inputs or Assumptions: Finally, the quality of the model outputs depend on the quality of the input data and assumptions, and errors in inputs or incorrect assumptions will lead to inaccurate outputs.
19. Models used Incorrectly or Inappropriately: The model may also be used incorrectly or inappropriately.
20. Misapplication of a Sound Model: Even a fundamentally sound model producing accurate outputs consistent with the design objectives of the model may exhibit high model risk if it is misapplied or misused.
21. Inappropriate Simplification of Real-World Events: Models by their nature are simplifications of reality, and real-world events may prove those simplifications inappropriate.
22. Invalid Usage of the Model: This is even more of a concern if a model is used outside of the environment for which it was designed.
23. Causes of Inappropriate Model Usage: Banks may do this intentionally as they apply existing models to new products or markets, or inadvertently as the market condition or the customer behavior changes.
24. Consistency in the Model Usage: Decision makers need to understand the limitations of a model and avoid using it in ways that are not consistent with the original intent.
25. Usage under Model Limitations Conditions: Limitations come in part from weaknesses in the models due to its various shortcomings, approximations, and uncertainties.
26. Usage out of Model Scope: Limitations are also a consequence of assumptions underlying a model that may restrict the scope to a limited set of circumstances and situations.
27. Management of the Model Risk: Model risk should be managed like other types of risk.
28. Identifying Risk Sources and Magnitude: Banks should identify the sources of risk and assess the magnitude.

29. Impact of the Model Risk: Model risk increases with greater model complexity, high uncertainties about inputs and assumptions, broader use, and larger potential impact.
30. Aggregate and Decomposed Model Risk: Banks should treat risk from individual trades and in the aggregate.
31. Causes of Aggregate Model Risk: Aggregate model risk is affected by interactions and dependencies among the models; reliance on common assumptions, data, or methodologies; and any other factors that could adversely affect several models and their outputs at the same time.
32. Model Risk Origin and Impact: With an understanding of the source and the magnitude of model risk in place, the next step is to manage it properly.
33. Model Risk Management Guiding Principles: A guiding principle for managing model risk is *effective challenge* of models, that is, critical analysis by objective informed parties who can identify model limitations and assumptions and produce appropriate changes.
34. Effective Challenge of a Model: Effective challenge depends on a combination of incentives, competence, and influence.
35. Incentivizing the Effective Model Challenge: Incentives to produce effective challenges to models are stringer when there is a greater separation of that challenge from the model development process and when that challenge is supported by well-designed compensations practices and corporate culture.
36. Competence helps Effective Model Challenge: Competence is a key to effective model challenge since technical knowledge and modeling skills are necessary to conduct appropriate analysis and critique.
37. Influence for an Effective Critique: Finally, the challenge may fail to be effective without an influence to ensure that actions are taken to address model issues.
38. Characteristics Aiding Effectiveness of Critique: Such influence comes from a combination of explicit authority, stature within the organization, and commitment and support from the higher levels of management.
39. Need to Alleviate Model Risk: Even with skilled modeling and robust validation, model risk can still not be eliminated, so other tools must be used to manage model risks effectively.

40. Strategies for Alleviating Model Risk: Among these are: establishing limits on model use; monitoring model performance; adjusting or revising models over time; and supplementing model results with other analysis and implementation.
41. Balanced and Informed Model Conservatism: Informed conservatism, in either the inputs or in the design of a model, or though explicit adjustments to the outputs, can be an effective tool, though not an excuse for improving models.
42. Impact Materiality of a Model: As is generally the case with other risks, materiality is an important consideration in model risk management.
43. Low Model Impact Materiality: In some banks, if the use of models is less pervasive, and has less of an impact on their financial condition, then those banks may not need as complex an approach to model risk management in order to meet supervisory expectations.
44. Managing High Model Materiality Impact: However, where models and model outputs have a material impact on the business decisions, including decisions related to risk management and capital and liquidity planning, and where model failures would have a particularly harmful impact on the bank's financial condition, a bank's model risk management framework should be more extensive and rigorous.
45. Robust Model Development and Use: Model risk management begins with robust model development, implementation, and use.
46. Soundness of the Model Validation Process: Another essential element is the sound model validation process.
47. Importance of Model Governance Function: The third element is governance, which sets an effective framework with defined roles and responsibilities for clear communication of model limitations and assumptions, as well as the authority to restrict model use. The following sections of this chapter cover each of these elements.

Model Development, Implementation, and Use

1. Structured and Competent Model Development: Model risk management should include disciplined and knowledgeable model implementation processes that are consistent with the situation and the goals of model users and with bank policy.
2. Individualized/Customized Model Development Process: Model development is not a straightforward or a routine technical process.
3. Model Developers' Experience and Judgement: The experience and judgement of the developers, as much as their technical knowledge, greatly influence the selection of appropriate inputs and the processing components.
4. Impact of Model Developers' Training/Judgement: The training and the experience of the developers exercising such judgement affects the extent of model risk.
5. Multi-disciplinary Nature of Model Development: Moreover, the model exercise is often a multi-disciplinary activity drawing on economic, finance, statistics, mathematics, and other fields.
6. Model Tailored for Business Use: Models are deployed for use in real-world markets and events, and should therefore be tailored for specific applications and informed by business users.
7. Subjective Nature of Model Development: In addition, a considerable amount of subjective judgement is exercised at various stages of model development, implementation, use, and validation.
8. Subjectivity Impact on Soundness and Comprehensiveness: It is important for the decision makers to recognize that this subjectiveness elevates the importance of sound and comprehensive model risk management process.
9. Reliance on Vendor Models - #1: Smaller banks that rely on vendor models may be able to satisfy the standards in this guidance without an in-house staff of technical, quantitative model developers.
10. Reliance on Vendor Models - #2: However, even if a bank relies on vendor models for basic model development, it should still choose the particular models and variables appropriate for its size, scale, and lines of business and ensure that the models are appropriate for their intended use.

Model Development and Implementation

1. Model Development Aligned with its Use: An effective development process begins with a clear statement of purpose to ensure that the model development is aligned with its intended use.
2. Documenting the Design, Theory, and Usage: The design, the theory, and the logic underlying the model should be well-documented and generally supported by published research and sound industry practices.
3. Details of Model Methodologies and Components: The model methodologies and the processing components that implement the theory, including the mathematical specifications and the numerical techniques and approximations, should be explained in detail with particular attention paid to merits and de-merits.
4. Model Governance and Statistical Validity: Developers should ensure that the components work as intended, are appropriate for the intended business purpose, are conceptually sound, and are mathematically and statistically valid.
5. Comparison against Alternate Theories/Approaches: Comparisons with alternate theories/approaches is a fundamental component of a sound modeling process.
6. Assessing Data Quality and Relevance: The data and other information used to develop a model are of critical importance; there should be rigorous assessment of data quality and relevance, and appropriate documentation.
7. Data Compatibility with the Methodology: Developers should be able to demonstrate the such data and information are suitable for the model and that they are consistent with the theory behind the approach and with the chosen methodology.
8. Data Proxies - Identification, Justification, Documentation: If data proxies are used, they should be carefully identified, justified, and documented.
9. Documentation of the Data Adjustment: If the data and the information are not representative of the bank's portfolio or other characteristics, or if assumptions are made to adjust the data

or other information, these factors should be properly tracked and analyzed so that the users are aware of potential limitations.

10. Special Attention Paid to External Data: This is particularly important to external data and information – from a vendor or outside party – especially as they relate to new products, instruments, or activities.
11. Model Unit and Integration Testing: An integral part of model development is testing, in which the various components of the model and its overall functioning are evaluated to determine whether the model is performing as intended.
12. Checking the Model's Behavioral Aspects: Model testing includes checking the model's accuracy, demonstrating the model is robust and stable, assessing potential limitations, and evaluating the model's behavior over a range of input values.
13. Impact of Assumptions and Problem Situations: It should also assess the impact of assumptions and identify situations where the model performs poorly or becomes unreliable.
14. Test Coverage Products, Markets, Scenarios: Testing should be applied to actual circumstances under a variety of market conditions, including scenarios that are outside the range of ordinary expectations, and should encompass the variety of products or approaches for which the model is intended.
15. Model Effectiveness at the Boundaries: Extreme values of inputs should be evaluated to identify any boundaries of model effectiveness.
16. Impact Assessment on Downstream Models: The impact of model results on other models that rely on the results as inputs should also be evaluated.
17. Compilation of Test Plans and Results: Included in the testing activities should be the purpose, the design, and the execution of the test plans, summary results with commentary and evaluation, and detailed analysis of informative samples.
18. Documentation of the Tests Covered: Testing activities should be appropriately documented.
19. Model Testing and Analysis: The nature of testing and analysis will depend on the type of the model and will be judged by different criteria depending on the context.
20. Example - Statistical Model Hypothesis Tests: For example, the appropriate statistical tests depend on the specific distributional assumptions and the purpose of the model.

21. Unambiguously Identifying TRUE/FASLE Hypothesis: Furthermore, in many cases, statistical tests cannot unambiguously reject FALSE hypothesis and accept TRUE ones based on sample information.
22. Test Suite Strengths and Weaknesses: Different tests have different strengths and weaknesses under different conditions.
23. Need for Varieties of Tests: Any single tests is barely sufficient, so banks should apply a variety of tests to develop a sound model.
24. Soundness of Qualitative/Judgmental Aspects: Banks should ensure that the development of the more qualitative and judgmental aspects of their models is also sound.
25. Example Tweaks of Statistical Outputs: In some cases, banks may take a statistical output from a model and modify it with judgmental or qualitative adjustments as part of model development.
26. Systematic Incorporation of Input Adjustments: While such practices may be appropriate, banks should ensure that any such adjustments made as part of the development process are conducted in an appropriate and systematic manner, and are well documented.
27. Model Incorporation inside Data Systems: Models are typically embedded in large information systems that manage the flow of data from various sources into the model and handle the aggregation and the reporting of the model outcomes.
28. System Integration of Model Results: Model calculations should be properly coordinated with the capabilities and the requirements of the information systems.
29. Model Integration inside Information Systems: Sound model risk management depends on substantial investment on supporting systems to ensure data and reporting integrity, together with controls and testing to ensure proper implementation of models, effective systems integration, and appropriate use.

Model Use

1. Model Use over Real-World: Model use provides additional opportunity to test whether a model is functioning effectively and to assess its performance over time as conditions and model applications change.
2. Feedback from the Impacted Community: It can serve as a source of productive insight and feedback from a knowledgeable internal constituency with string interest in having models that function well and reflect economic and business realities.
3. Insights from the Model Users: Model users can provide valuable business insight during the development process.
4. Critiques Raised by Business Users: In addition, business managers affected by model outcomes ay question the methods or outcomes underlying the models, particularly if the managers are significantly affected by, and do not agree with, the outcome.
5. Feedback into the Design Process: Such questioning can be healthy if it is constructive and causes model developers to explain and justify the assumptions and the design of the models.
6. Weak Challenges from Model Users: However, challenges from the model users may be weak if the models do not materially affect their results, if the resulting changes in models are perceived to have a adverse effect on the business line, or if the change in general is regarded as expensive or difficult.
7. Incomplete Usage from Business Users: User challenges also tend not to be comprehensive because they focus on the aspects of models that have most direct impact on the user's measured business performance or compensation, and thus may ignore other elements and applications of the models.
8. Asymmetry in the Users' Challenge: Finally, such challenges tend to be asymmetric, because users are less likely to challenge an outcome that results in an advantage for them.
9. Appearance of Low Model Risk: Indeed, users may incorrectly believe that the model risk is low simply because outcomes from model-based decisions appear favorable to the institution.
10. Examining the User's Model Focus: Thus, the nature and the motivation behind the model user's inputs should be evaluated carefully, and banks should also solicit constructive suggestions and criticisms from sources independent of the line of business using the model.
11. Business Decisions' Role on Model Risk Management: Reports used for business decision making play a critical role in model risk management.

12. Accounting for Varying Model Interpretations: Such reports should be clear and comprehensible and take into account the fact that decision makers and modelers quite often come from different backgrounds and may interpret the contents in different ways.
13. Reports Providing Range Based Accuracy: Reports that provide a range of estimates for different input value scenarios and assumptions can give decision makers important indications of the model's accuracy, robustness, and stability, as well as information on model limitations.
14. Accounting for Model Inaccuracy and Uncertainty: An understanding of model uncertainty and inaccuracy and a demonstration that the bank is accounting for them appropriately are important outcomes of model development, implementation, and use.
15. Models as Imperfect Representations of Reality: Because they are by definition imperfect representations of reality, all models have some degree of uncertainty and inaccuracy.
16. Quantifying Inaccuracy and Uncertainty: These can sometimes be quantified, for example, by an assessment of the potential impact of factors that are unobservable or not fully incorporated into the model, or by the confidence interval around a statistical model's point estimate.
17. Output as a Range: Indeed, using a range of output, rather than a simple point estimate, can be a useful way to signal model uncertainty and avoid spurious precision.
18. Assessing Inaccuracy and Uncertainty Qualitatively: At other times only a qualitative assessment of the model uncertainty and inaccuracy is possible.
19. Alternate Ways of Accounting for Uncertainty: In either case, it can be prudent for banks to account for model uncertainty by explicitly model inputs or calculations to produce more severe or adverse model outputs in the interest of conservatism.
20. Conditional Usage and Output Augmentation: Accounting for model uncertainty can also provide conservative judgements to model output, placing less emphasis on that model's output, or ensuring that the model is used only when supplemented by other models or approaches.
21. GAAP Compliance of Financial Statements: To the extent that models are used to generate amounts included in public financial statements, any adjustment for model uncertainty must comply with Generally Accepted Accounting Principles.

22. Blind Application of Conservatism: While conservative use of models is prudent in general, banks should be careful in applying conservatism broadly or claiming to make more conservative adjustments or add-ons to address model risk, because the impact of such conservatism in complex models may not be obvious or intuitive.
23. Uniformity of Conservatism across Models: Aspects that may appear conservative under one model may not be truly conservative compared with alternative models.
24. Example - Conservatism under Model Mis-specification: For example, simply picking an extreme point under a given modeled distribution may not be conservative if the distribution was misestimated or mis-specified in the first place.
25. Reduction in Conservatism over Time: Furthermore, initially conservative assumptions may not remain conservative over time.
26. Definition and Measurement of Conservatism: Therefore, banks should justify and substantiate claims that model outputs are conservative, with a definition and measurement of conservatism that is communicated to model users.
27. Sensitivity Analysis and Stress Testing: In some cases, sensitivity analysis and other types of stress testing can be used to demonstrate that a model is indeed conservative.
28. Model Risk Capital Buffer Cushion: Another way in which banks may choose to be conservative is to hold an additional cushion of capital to protect against losses associated with model risk.
29. Conservatism not an Alternative to Bad Models: However, conservatism can become an impediment to proper model development and application if it is seen as a solution that dissuades the bank from making an effort to improve the model; in addition, excess conservatism can lead the model users to discount the model outputs.
30. Robust Model Development and Implementation: As this section has explained, robust model development, implementation, and use are important to model risk management.
31. Limitations of Model Understanding/Acceptance: But it is not enough for the model developers and users to simply understand and accept the model.
32. Importance of the Model Validation Process: Because the model risk is ultimately borne by the bank as a whole, the bank should objectively assess model risk and the associated costs and benefits of using a sound model validation process.

Model Validation

1. Formal Definition of Model Validation: Model validation is a set of activities and processes intended to verify that the model is performing as expected, in line with the design objectives and business uses.
2. Ensuring the Soundness of Models: Effective validation helps ensure that the models are sound.
3. Impact, Limitations, and Assumptions: It also identifies potential limitations and assumptions, and assesses their possible impact.
4. Aspects of Effective Model Validation: As with other aspects of effective challenge, model validation should be performed by staff members with appropriate incentives, competence, and influence.
5. Model Components Requiring Validation: All model components, including input processing, and reporting, should be subject to validation; this applies equally to models developed in-house as well as to those purchased from or developed by vendors or consultants.
6. Rigor and Sophistication of Model Validation: The rigor and sophistication of validation should be commensurate with the Bank's overall use of models, the complexity and the materiality of the models, and the complexity of the bank's operations.
7. Independence Needed for Model Validation: Validation involves a degree of independence for model development and use.
8. Validators are NOT Model Developers/Users: Validation should be done by people who are responsible for model development or use, and do not have a stake in whether a model is determined to be valid.
9. Independence aiding Model Validation Goals: Independence is not an end in itself, but rather helps ensure that incentives are aligned with the goals of model validation.

10. Assessing the Impact of Model Validation: While independence may be supported by separating the reporting lines, it should be judged by actions and outcomes, since there may be other ways to ensure objectivity and bias.
11. Reviewing Validations done by Developers/Users: As a practical matter, some validation work may be done most effectively by model developers and users; it is essential, however, that such validation work be subject to critical review by an independent party, who should conduct additional activities to ensure proper validation.
12. Critical Review as Part of the Process: Overall, the quality of the process is judged by the manner in which models are subject to critical review.
13. Documentation, Issues Identification, and Actions Taken: This could be determined by evaluating the extent and the clarity of the documentation, the issues identified by objective parties, and the actions taken by the management to address model issues.
14. Incentives Based Model Validation Practice: In addition to independence, banks can support appropriate incentives in validation through compensation practices and performance evaluation standards that are tied directly to the quality of model validation and the degree of critical, unbiased review.
15. Corporate Culture of Critical Challenges: In addition, corporate culture plays an important role if it establishes support for critical thinking and encourages questioning and challenging of decisions.
16. Validation needs Knowledge, Skills, and Expertise: Skills doing the validation should have the requisite knowledge, skills, and expertise.
17. Technical Expertise to Handle Model Complexity: A high level of technical expertise may be needed because of the complexity of many models, both in structure and in application.
18. Familiarity with the Model Usage: These staff should also have a significant degree of familiarity with the line of business using the model and the model's intended use.
19. Reliance on the Model Developer: A model developer is an important source of information, but cannot be relied on as an objective or as a sole source on which to base an assessment of the model quality.

20. Explicit Authority to Challenge Models: Staff conducting validation work should have explicit authority to challenge developers and users and to elevate their findings, including issues and deficiencies.
21. Influence of the Model Validation Unit: The individual or the unit to whom the staff report should have sufficient influence ort stature within the bank to ensure that any issues and deficiencies are appropriately addressed in a timely and appropriate manner.
22. Mechanism to Establish the Influence: Such influence can be reflected in reporting lines, title, rank, or designated responsibilities.
23. Influence Elevation via Effective Calls: Influence may also be demonstrated by a pattern of actual instances in which models, or the use of models, have been appropriately changed as a result of validation.
24. Validation before Initial Use: The range and the rigor of validation activities conducted before the first use of a model should be in line with the potential risk presented by the use of the model.
25. Deficiencies Uncovered during Initial Validation: If significant deficiencies are noted as a result of the validation process, use of the model should not be allowed, or should be permitted only under very tight constraints until those issues are resolved.
26. Handling Highly Significant Model Deficiencies: If the deficiencies are too severe to be addressed within the model's framework, the model should be rejected.
27. Infeasibility of Comprehensive Initial Validation: If it is not feasible to conduct necessary validation activities prior to model use because of data paucity or other limitations, the fact should be documented and communicated in reports to users, senior management, and other relevant parties.
28. Limited Validations - Use of Compensating Controls: In such cases, the uncertainty about the results that the model produces should be mitigated by other compensating controls.
29. New/Existing Models and Applications: This is particularly applicable to new models, and to the use of existing models in new applications.
30. Validation on a Continual Basis: Validation activities should continue on an ongoing basis after a model goes into use, to track known model limitations and to identify new ones.

31. Checks during Benign Financial Conditions: Validation is an important check on model use during periods of benign economic and financial conditions, when estimates of risk and potential loss can become overly optimistic, and when the data at hand does not reflect the more stressed conditions.
32. Tracking of New Model Limitations: Ongoing validation activities help to ensure that changes in markets, products, exposures, activities, clients, and business practices do not result in new model limitations.
33. New Limitations Example - Under-writing Changes: For example, if the credit risk models do not incorporate under-writing changes in a timely manner, flawed and costly business decisions could be made before deterioration in model performance becomes apparent.
34. Periodic Review of Model Inventory: Banks should conduct a periodic review – at least annually but more frequently if warranted – of each model to determine whether it is working as intended and if existing validation activities are sufficient.
35. Actions Resulting from Periodic Reviews: Such a determination could simply affirm previous validation work, suggest updates to previous validation activities, or call for additional validation activities.
36. Material Changes to Model Details: Material changes to models should also be subject to validation.
37. Periodic Full Review of Validation: It is generally good practice for banks to ensure that all models undergo full validation process, as described in the following section, at some fixed interval, including updated documentation of all activities.
38. Advantages of Effective Model Validation: Effective model validation helps reduce model risk by reducing model errors and ensuring actions and appropriate use.
39. Criteria Based Model Reliability Assessment: It also provides an assessment of the reliability of a given model based on its underlying assumptions, theory, and methods.
40. Model Risk Origin and Size: In this way, it provides information about the source and the extent of model risk.
41. Timely Detection of Performance Degradation: Validation can also reveal deterioration of model performance over time and can set thresholds for acceptable levels of error, through analysis of the distribution of outcomes around the expected or the predicted values.

42. Consistent Under-performance of a Model: If outcomes fall consistently outside this acceptable range, the models should then be re-developed.

Key Elements of Comprehensive Validation

1. Components of Effective Validation Framework: An effective validation framework should include the following three core elements.
2. Evaluation of the Model's Conceptual Soundness: Evaluation of conceptual soundness, including developmental evidence.
3. Ongoing Process Verification and Benchmarking: Ongoing monitoring, including process verification and benchmarking.
4. Statistical Hypotheses Tests Outcomes Analysis: Outcomes analysis, including back-testing.

Evaluation of Conceptual Soundness

1. Assessing Design and Construction Quality: This element involves assessing the quality of model design and construction.
2. Reviewing the Documentation and Evidence: It entails reviewing the documentation and the empirical evidence supporting the methods used and the variable selected for the model.
3. Documentation/Testing Limitations and Assumptions: Documentation and testing should convey an understanding of the model limitations and assumptions.

4. Published Research and Industry Practice: Validation should ensure that the judgement exercised in model design and construction is well informed, carefully considered, consistent with published research and with sound industry practice.
5. Reviewing Developmental Evidence before Use: Developmental evidence should be reviewed before a model goes into use and also as part of ongoing validation process, in particular whenever there is a material change in the model.
6. Documentation Supporting all Model Choices: A sound development process will produce documented evidence in support of all model choices, including the overall theoretical construction, key assumptions, data, and specific mathematical calculations, as mentioned earlier.
7. Critical Analysis, Evaluation, and Testing: As part of model validation, these model aspects should be subject to critical analysis by both evaluating the quality and the extent of developmental evidence and conducting additional analysis and testing as necessary.
8. Comparison with Alternate Theories and Approaches: Comparison with alternate theories and approaches should be included.
9. Assumptions, Inputs, Outputs, and Limitations: Key assumptions and the choice of variables should be assessed, with analysis on their impact on model outputs, and particular focus on any potential limitations.
10. Data Relevance to Usage Conditions: The relevance of the data used to build the model should be evaluated to ensure that it is reasonably representative of the bank's portfolio or market conditions, depending on the type of the model.
11. External Data or New Usage: This is an especially important exercise when the bank uses external data or the model is used for new products or services.
12. Sensitivity Analysis in Model Validation: Where appropriate to the particular model, banks should employ sensitivity analysis in model development and validation to check the impact of small inputs and parameter values on model outputs to make sure that they fall within the expected range.
13. High Model Sensitivity: Unexpected large changes in outputs to small changes in inputs can indicate an unstable model.

14. Sensitivity Analysis to Multiple Inputs: Varying several inputs simultaneously as part of the sensitivity analysis can provide evidence of unexpected interactions, particularly if the interactions are complex and not intuitively clear.
15. Stress Testing to Extreme Inputs: Banks benefit from conducting model stress testing to check performance over a wide range of inputs and parameter values, including extreme values, to verify that the model is robust
16. Establishing Boundaries of Model Performance: Such testing helps establish boundaries of model performance by identifying the acceptable range of inputs as well as conditions under which the model may become unstable or inaccurate.
17. Operationalizing Sensitivity and Scenario Analysis: Management should have a clear plan for using sensitivity analysis and other quantitative/qualitative testing.
18. Handling Unstable Input Parameter Range: If testing indicates that the model may be inaccurate or unstable in some circumstances, management should consider modifying certain model properties, putting less reliance on its outputs, placing limits on model use, or developing a new approach.
19. Evaluation of Qualitative Judgement Inputs: Qualitative information and judgement used in model development should be evaluated, including the logic, the judgement, and the types of information used, to establish the conceptual soundness of the model and to set appropriate conditions for its use.
20. Systematic Approach of the Qualitative Judgments: The validation process should ensure that qualitative and judgmental processes are conducted in an appropriate and systematic manner, are well supported, and are documented.

References

- Federal Reserve (2011): Supervisory Guidance on Model Risk Management