Withdrawal Delay Cohort Method Under VM-21

By Benjamin Buttin, Matthias Kullowatz, Zi Xiang Low and Zohair Motiwalla

In early December 2017, the National Association of Insurance Commissioners (NAIC) released proposed revisions to the existing U.S. variable annuity statutory framework. These revisions were promulgated as redline updates to the existing Actuarial Guideline 43 (AG 43) and Risk Based Capital C-3 Phase II instructions, paving the way for VM-21 of the *Statutory Valuation Manual* (VM), "Requirements for Principle-Based Reserves for Variable Annuities." After an exposure period in early 2018 to allow for comments from industry participants, regulators and interested parties, the Variable Annuity Issues (E) Working Group of the NAIC adopted almost all of the recommended changes outlined in the redline instructions.

While these revisions have been broadly agreed upon by the NAIC, a final set of regulatory instructions for VM-21 is still pending, with the responsibility assigned to the VM-21 Report Drafting Group. New updated redline instructions are exposed publicly on a piecemeal basis, inviting comments and feedback from practitioners and interested parties.[1] The working expectation is that the final version of VM-21 will be formally adopted at the NAIC Summer Meeting in August 2019 for a Jan. 1, 2020, effective date. Under the new VM-21 framework, the Aggregate Reserve is now the sum of the conditional tail expectation 70 amount (CTE Amount) and the Additional Standard Projection Amount, where the latter term is determined using the Standard Projection.

The VM-21 Standard Projection is essentially a complete overhaul of the existing AG 43 Standard Scenario framework. It can be calculated using either the company-specific market path (CSMP) method or the conditional tail expectation with prescribed assumptions (CTEPA) method. The CSMP method uses at least 40 prescribed economic scenarios, while the CTEPA method uses the same economic scenarios as the CTE Amount calculation.

One of the more challenging and important components of the Standard Projection is the withdrawal delay cohort method (WDCM), which is a prescribed approach for determining the timing of policyholder election for policies with either hybrid guaranteed minimum income benefits (GMIB)[2] or guaranteed minimum withdrawal benefits (GMWB). This article discusses practical considerations when implementing the WDCM.

WDCM PROCESS

The WDCM applies in both the CSMP method and the CTEPA method. To be in scope for the WDCM, policies must be either nonconforming (meaning they have taken a withdrawal in the policy year occurring coincident with the valuation date, and this withdrawal was in excess of the GMWB's guaranteed annual withdrawal amount or the GMIB's dollar-for-dollar maximum withdrawal amount) or nonwithdrawers (meaning that they have not started taking withdrawals).

Under the existing AG 43 framework, the Standard Scenario assumes that the exercise of any living benefits such as GMIBs or GMWBs occurs at the earliest available opportunity that is consistent with contractual provisions.

In contrast, the WDCM under VM-21 defines a prescriptive process for determining a distribution of possible election cohorts for each policy in scope, each with its own weight. The cohorts simulate each potential age of starting systematic withdrawals. In order to determine the election distribution, the guaranteed actuarial present value (GAPV) concept, as prescribed under VM-21, is used to calculate the prospective withdrawal value of the rider to the policyholder at each potential individual withdrawal age.

The main steps in the WDCM are outlined below:

- For each potential initial withdrawal age (starting from issue), compute the GAPV assuming the policyholder elects to take withdrawals at that age. This will produce a set of GAPVs.
- Apply certain prescribed transformations and normalizations to this set of GAPVs to develop a from-issue cumulative distribution function (CDF), reflecting shocks as necessary. [3] This CDF defines a specific weight for the withdrawal cohort corresponding to each initial withdrawal age from issue.
- A "never withdraw" cohort is also defined, whose weight varies by rider type and tax status.
- Given a valuation date, any withdrawal cohorts corresponding to initial withdrawal ages occurring prior to that date are discarded and the remaining weights are rescaled to produce a rescaled CDF.

The key drivers in this process are those that underlie the GAPV calculation, namely the rider benefit base mechanics, the payout rate for the GMWBs and/or hybrid GMIBs under consideration, the prescribed Standard Projection mortality and the discount rate (3 percent). The most recent redline instructions stipulate that the CDF is calculated once for a set of policies with the same combination of issue age, rider type and tax status. For the purposes of this article, we refer to this combination as the WDCM cell key. In practice, there may be legitimate reasons to expand the WDCM cell key definition. For example, gender is a key item that should also be considered (because mortality rates will vary by gender). Moreover, the payout rate may vary by joint life status or rider generation.

Theoretically, policies with the same WDCM cell key should produce the same from-issue CDF even if their benefit bases on the valuation date are different, because the associated GAPVs should simply scale and the weights would renormalize to the same values. One could even calculate the CDF using an arbitrary (but nonzero) benefit base amount. Accordingly, for existing policies, the calculation of the from-issue CDF is intended to be a one-time process. Once calculated for a given WDCM cell key, the weights are fixed and do not need to be recomputed in the future.[4] The practitioner need only compute new weights for new business issued that have different WDCM cell key combinations.

USING RANDOM SAMPLING TO MITIGATE COMPUTATIONAL BURDEN

While the WDCM process is theoretically very appealing, in practice the run-time associated with splitting the in-force file into many cohorts (some of which may be assigned very small weights) can be very challenging, particularly under the CTEPA method. The full WDCM cohort file record count is likely to be many times greater than that of the original in-force file.

The redline instructions provide some allowance for discarding additional cohorts to mitigate the computational burden, so long as this decision has been disclosed. The specific language indicates that individual withdrawal age cohorts may be discarded or a small number of withdrawal cohorts may be assigned to each contract via random sampling.

Discarding cohorts to relieve the computation burden without loss of accuracy (relative to results produced using the full WDCM cohort approach) requires practitioners to engage in some analysis and testing, ideally before VM-21 becomes effective.

As noted in the redline instructions, one possible route practitioners can take is to use a random draw to collapse all cohorts to a single cohort for each in-force policy. The process would involve using a robust random number generator to produce a random draw on the interval zero to one for each in-force policy. This value would be compared with the rescaled CDF produced by the WDCM process, thereby randomly selecting a future election time and modeling each in-force policy using a single cohort with that particular election time. The advantage to this approach is that the in-force file record count for the randomized run is the same as the pre-WDCM version (i.e., the original in-force file). For proof of principle, the practitioner should verify that the results produced using both the random sampling approach and the full WDCM cohort approach are not only similar, but that repeated random trials produce stable results. This test should be performed at the onset of adopting the random sampling approach and may also need to be carried out at future intervals (such as to support disclosure of the approach in the year-end actuarial memorandum). It should be noted that a number of companies already employ random sampling methods in their CTE Amount calculations.

STATISTICAL THEORY BEHIND RANDOM SAMPLING

In defense of the random sampling approach outlined above (in which a single delay cohort is randomly selected for each policy) we argue that the greatest present value of accumulated deficiencies (GPVAD) calculated by randomly sampling the election time for each in-force policy will converge to the true GPVAD within an economic scenario for large in-force sizes, where the true GPVAD is that which would be calculated by using the full WDCM cohort in-force file. We start by showing convergence of the policy-level accumulated product cash flows, and we expand that to the convergence of the GPVAD.

Probability theory suggests that when you sample values from a population, the ratio between the sample standard deviation and the sample sum shrinks as the sample size increases. The sample standard deviation here can be thought of as an error, the discrepancy between our GPVAD estimate and the true GPVAD. As such, even though larger in-force sizes will generally lead to larger errors, the errors will become smaller as a proportion of total GPVAD.

This theory extends naturally to WDCM cohort sampling—which is effectively a form of stratified sampling—where exactly one outcome is randomly selected for each policy. We first conceptualize the effect using the policy-level accumulated product cash flows. Each policy has a theoretical variance of possible accumulated product cash flow values based on the randomness of which WDCM cohort is sampled. Because WDCM cohorts are sampled independently for each policy, the variance of the sum is equal to the sum of the variances, shown mathematically here:

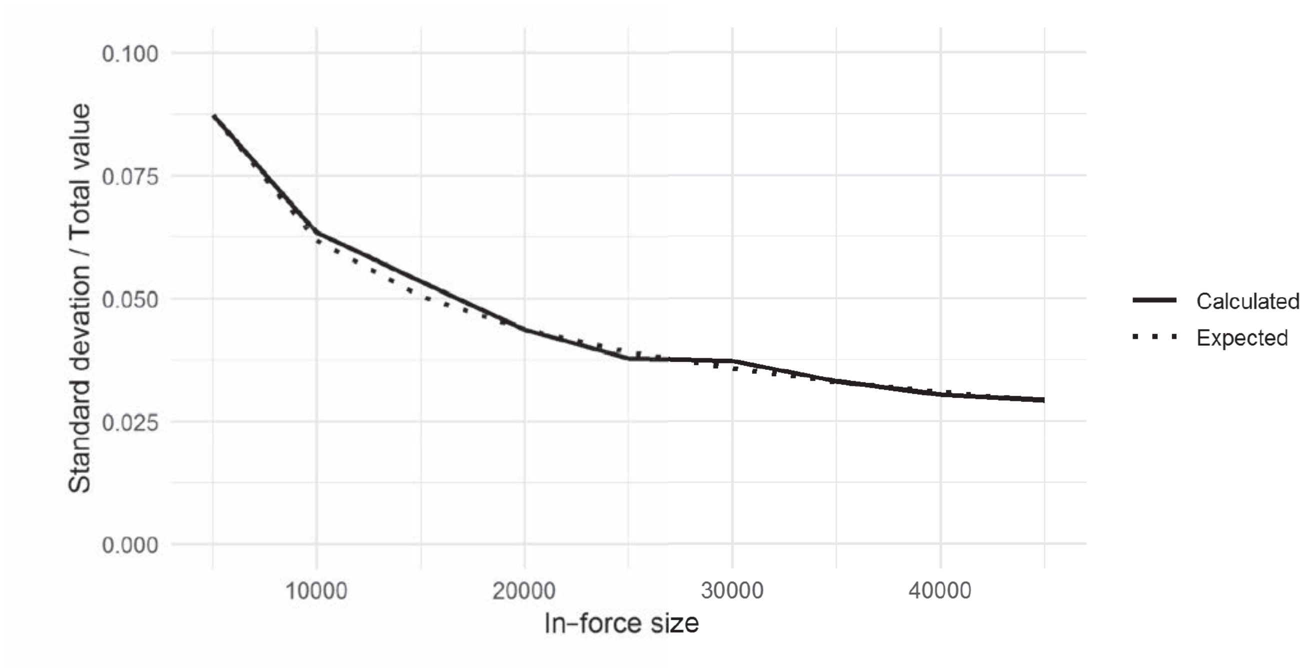
$$Var\left(\sum_{1}^{n} X_i\right) = \sum_{1}^{n} Var(X_i)$$

$$where\ X_i = sampled\ cash\ flow\ value\ for\ i^{th}\ policy\ and$$
$$n = in\text{-}\ force\ size$$

As such, the variance of the sum increases linearly with the in-force size, implying that the standard deviation of the sum increases at a rate *proportional to the square root* of the in-force size. In other words, the sum is growing at a linear rate, but the standard deviation, or error, is growing at the rate of the square root, which is much slower.

In order to illustrate this relationship, we started with nine sets of in-force files that contained samples of between 5,000 and 45,000 policies. Each of these in-force files contained policies that were cohorted under the prescribed full WDCM approach with accumulated product cash flow results pre-calculated

Figure 1
Ratio of Standard Deviation to Total Accumulated Product Cash Flows by In-Force Size



for each cohort. For each of these in-force files, we randomly sampled distinct sets of cohorts 1,000 times to generate a distribution of potential total accumulated product cash flows.

In Figure 1 (pg. 28), the solid line represents the ratio of the standard deviation of the random samples to the total accumulated product cash flows for each in-force file size, while the dotted line represents the ratio that we would expect to see if the square root principle held. The graph shown in Figure 1 explains the phenomenon near perfectly. In other words, the sample error—as measured by the sample standard deviation—will shrink at a rate proportional to the square root of the in-force size.

While the probability theory discussed in this article explains the variation for sums of policy-level cash flows quite well, it does not cover how convergence of a policy-level cash flow implies convergence of the GPVAD. Intuitively, the calculation of GPVAD implies additional aggregation, both within and across time steps, and aggregation generally leads to lower variances. For example, this concept of aggregation is used to diversify portfolios and reduce risk. We found that the relative error of GPVAD values across random samples was, in fact, lower than the relative error of policy-level cash flows for equally sized in-force blocks.[5]

FINAL THOUGHTS

In recognition of the potential run-time challenges posed by the WDCM for variable annuity statutory valuation requirements under the VM-21 Standard Projection, we expect that companies will be looking to incorporate innovative solutions to manage the computational burden. Random sampling offers one such solution—one that is allowed within the proposed framework.

A complete version of this article that also presents a WDCM case study comparing the random sampling approach with the prescribed full WDCM approach for a guaranteed living withdrawal benefit (GLWB) block of business can be found at the following website address: *http://www.milliman.com/insight/2019/The-Withdrawal-Delay-Cohort-under-VM-21/AG-43-The-case-for-random-sampling/*. Certain technical considerations for companies thinking of adopting the random sampling approach are also discussed. ■

Benjamin Buttin, ASA, MAAA, is an associate actuary at Milliman. He can be reached at *benjamin.buttin@milliman.com.*

Matthias Kullowatz, ASA, MAAA, is an associate actuary at Milliman. He can be reached at *matthias.kullowatz@milliman.com.*

Zi Xiang Low, FSA, FIA, MAAA, is an actuary at Milliman. He can be reached at *zixiang.low@milliman.com.*

Zohair Motiwalla, FSA, MAAA, is a principal and consulting actuary at Milliman. He can be reached at *zohair.motiwalla@milliman.com.*

ENDNOTES

[1] This article has been developed using the updated VM-21 redline that was exposed in early March 2019. The reader is cautioned that to the extent that the final version of the instructions is different from this redline, certain outcomes from this article may need to be revised.

[2] A hybrid GMIB policy is a policy with both guaranteed growth (such as with a rollup or doubler) and dollar-for-dollar partial withdrawal reductions in the GMIB benefit base.

[3] For applicable policies, these prescribed shocks correspond to the end of the rollup period and/or required minimum distributions after age 70 for qualified plans.

[4] Other than for the rescaling as the valuation date changes. Also, if there is a model correction/refinement that impacts the key drivers outlined above, then the CDFs need to be recalculated.

[5] One can find our case study on GPVAD stabilization in the complete version of this article, linked in the Final Thoughts section.