# Rebalance, Don't De-Risk: Lessons from 10,000 Simulated Retirements Under Market Regime Uncertainty

Filip Zaleski Wen Rong Lena Ho

NYU Stern School of Business

10th August 2026

**Abstract**

Retirees drawing down a fixed pool of savings face *sequence-of-returns risk*: the risk that the order in which good and bad market years arrive — not just their average — determines whether a portfolio lasts. We build a Monte Carlo simulation of a 30-year retirement in which annual equity, bond, and cash returns are generated by a two-state (calm/stressed) Markov regime-switching process, and use it to compare asset allocation, withdrawal, and rebalancing policies on their probability of depleting the portfolio and their sustainable withdrawal rate. Fitting the regime model to 154 years of US data (1872–2025) yields a persistent chain in which stressed years cluster and real equity returns fall from +12.6% to +0.6%. Across 10,000 replications the baseline 60/40 portfolio with a 4% fixed real withdrawal fails in 6.5% of retirements and supports a sustainable withdrawal rate of 4.34%, closely reproducing the classic "4% rule" benchmark. Annual rebalancing lowers the failure rate relative to buy-and-hold (6.5% versus 7.5%), while an age-based glide path does *not* help (6.9%) because it de-risks after the period of greatest danger has passed; percentage-of-balance rules eliminate depletion by construction but convert that risk into income volatility of roughly 24% of average spending.

## 1 Problem Description

Most classroom treatments of portfolio choice stop at a single-period question: given a menu of possible allocations, which one offers the best trade-off between expected return and risk. That question can largely be answered with closed-form formulas (mean-variance optimization, the Sharpe ratio) and does not require simulation at all.

The decision retirees actually face is harder, and it is inherently dynamic. A retiree does not hold a portfolio for one period — they draw down a fixed pool of savings over two or three decades, withdrawing money to live on while the remaining balance is still exposed to market risk. Two retirees who experience the exact same sequence of annual returns, just in a different order, can end up in very different places: a market crash in the first few years of retirement forces the sale of a large fraction of assets at depressed prices to fund that year's withdrawal, permanently shrinking the base that has to compound back up, whereas the same crash occurring twenty-five years in barely matters. This is *sequence-of-returns risk*, and it means the outcome for a retiree depends on the entire path of returns and regime conditions over their retirement, not merely on the average return. That path dependence is exactly the kind of feature that closed-form finance formulas cannot handle but that Monte Carlo simulation is built for: we can generate thousands of alternate 30-year "lives" for a retiree and observe what fraction of them go badly.

**Decision question.** For a retiree drawing down a portfolio over a 30-year retirement, which combination of asset allocation, withdrawal policy, and rebalancing rule minimizes the probability of depleting the portfolio, while sustaining an acceptable standard of living?

# 2 Model Formulation

## 2.1 System Description

A retiree begins with an initial real (inflation-adjusted) balance $B_0$, split between equities, bonds, and cash according to an allocation policy, and draws it down over $T = 30$ annual steps. Each year the retiree withdraws $W_t$, and the balance that remains earns that year's realized portfolio return before the next year begins. Depending on the scenario, the portfolio may be rebalanced back to its target weights at the end of the year.

Writing $\rho_t$ for the realized portfolio return in year $t$, the balance evolves as

$$B_t \;=\; \big(B_{t-1} - W_t\big)\big(1 + \rho_t\big), \qquad t = 1, \dots, T, \tag{1}$$

with the convention that the withdrawal is taken at the start of the year and the surviving balance is exposed to that year's market return. Depletion ("ruin") is the event that the balance cannot fund a withdrawal; the ruin time is

$$\tau \;=\; \min\{\, t \le T : B_{t-1} \le W_t \,\}, \tag{2}$$

with $\tau = \infty$ if the portfolio survives all $T$ years. Once ruined, the balance stays at zero for the remainder of the horizon.

## 2.2 Sources of Randomness

**Market regime.** Each year the market is in two market regimes, calm ($k = 1$) or stressed ($k = 2$). A year counts as *stressed* when its equity market was unusually turbulent — concretely, when the volatility of that year's monthly returns falls in the most volatile third of the historical record; the classification rule is stated precisely in Section 2.4. The regime $R_t$ evolves as a discrete-time Markov chain with a stationary $2 \times 2$ transition matrix $P$ estimated from historical data:

$$\Pr\big(R_t = j \mid R_{t-1} = i,\; R_{t-2}, \dots\big) \;=\; \Pr\big(R_t = j \mid R_{t-1} = i\big) \;=\; P_{ij}. \tag{3}$$

Each simulated retirement has to start in some regime, and choosing one arbitrarily would bias the results — starting every path in the calm state, for example, would understate early-retirement risk. We therefore draw $R_0$ from the chain's *stationary distribution* $\pi$: the long-run share of years the chain spends in each regime, given by the solution of $\pi P = \pi$ with $\sum_k \pi_k = 1$. Starting from $\pi$ means a simulated retiree is as likely to begin in a stressed market as history suggests they would be.

**Asset returns.** Conditional on the active regime, that year's real annual returns on the three assets are drawn jointly from a regime-specific multivariate normal:

$$\mathbf{r}_t \;=\; \big(r_t^{\text{eq}},\, r_t^{\text{bd}},\, r_t^{\text{csh}}\big)^\top \;\Big|\; R_t = k \;\;\sim\;\; \mathcal{N}\big(\boldsymbol{\mu}_k,\, \Sigma_k\big). \tag{4}$$

Drawing the three assets *jointly* rather than separately preserves their cross-asset correlations. In practice we generate a standard normal vector $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, I_3)$ and set $\mathbf{r}_t = \boldsymbol{\mu}_k + L_k \mathbf{z}_t$, where $L_k$ is the Cholesky factor of $\Sigma_k$. Given portfolio weights $\mathbf{w}_t$, the portfolio return in Equation (1) is $\rho_t = \mathbf{w}_t^\top \mathbf{r}_t$.

Because the return-generating process is a *mixture* of two normal distributions rather than a single one, the unconditional (long-run) distribution of returns is fat-tailed relative to a plain multivariate normal model.

## 2.3 Modeling Assumptions

- Two market regimes (calm, stressed), each with a fixed regime-specific mean/covariance; the regime persists or switches each year according to a stationary transition matrix (no further structural breaks).
- All quantities — returns, balances, withdrawals — are modeled in *real* (inflation-adjusted) terms, so no separate inflation adjustment is needed for withdrawals.
- Fixed 30-year retirement horizon, annual time steps, single initial lump-sum balance (no further contributions).
- No taxes, fees, or transaction costs (discussed as a limitation in Section 5).
- Regime and within-regime parameters are estimated once from the full historical sample and held fixed (no time-varying parameter estimation).

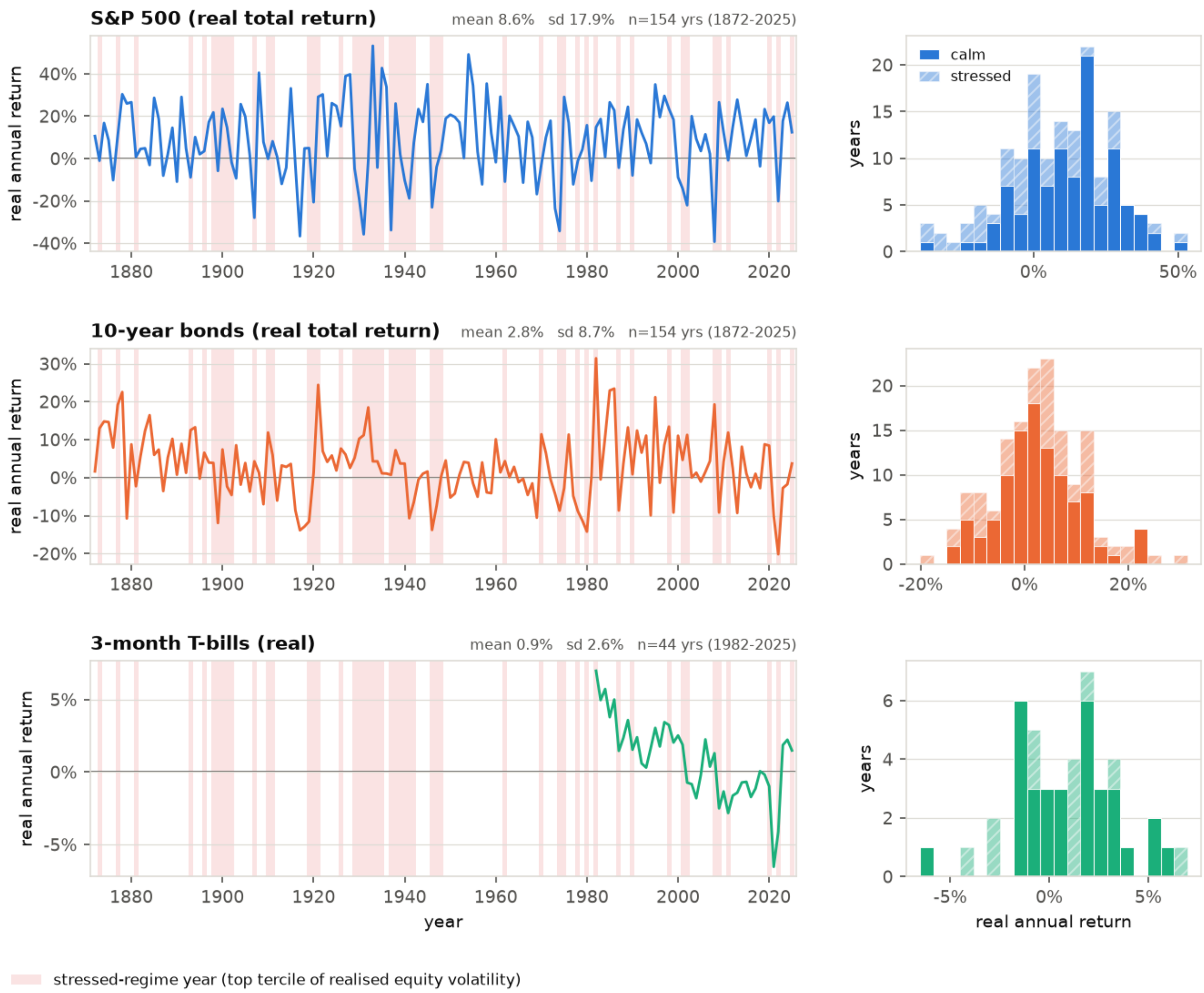
## 2.4 Data and Inputs

Two historical data sources are used to estimate the model's inputs:

- **Robert Shiller's long-run dataset** (`ie_data.xls`, Yale): monthly S&P 500 price, dividends, and CPI from January 1871 to present, plus constructed 10-year bond total-return series. The real, dividend-reinvested S&P 500 total-return index and the real 10-year bond total-return index are both used directly, for the equity and bond legs respectively.
- **FRED 3-Month Treasury yield** (`DGS3MO.csv`): daily 3-month T-bill yields from September 1981 to present, used as the cash leg. This series is a *yield level*, not a return index, so it is converted to a monthly return and deflated by CPI to place it on the same real basis as the other two. It is considerably shorter than the equity/bond history, which is noted as a limitation.

Monthly series are compounded to annual real returns, December to December, so that all three assets are measured over identical windows. Figure 1 shows the resulting series and their distributions. Over 1872–2025 real equity returns average 8.6% (s.d. 17.9%) and real bond returns 2.8% (s.d. 8.7%); real T-bill returns average 0.9% (s.d. 2.7%) over the shorter 1982–2025 window.

Figure 1: Time series and histograms of real annual returns for the S&P 500, 10-year bonds, and 3-month T-bills (full sample), with stressed-regime years shaded.




**Regime-fitting procedure.** Market regimes are not directly observed, so we construct calm and stressed labels from historical equity volatility. We use a two-stage estimator: classify every year, then count transitions between the resulting labels.

*Stage 1 — classification.* For each year $y$ we compute the *realized volatility* $s_y$, defined as the standard deviation of that year's twelve monthly real equity returns. This measures how turbulent the year was, as opposed to whether it happened to end up or down. A year is labelled stressed if that volatility lands in the top third of the historical distribution:

$$R_y = \begin{cases} \text{stressed} & \text{if } s_y > q_{2/3}, \\ \text{calm} & \text{otherwise,} \end{cases} \qquad q_{2/3} = \text{the 67th percentile of } \{s_y\}_{1872}^{2025} = 3.56\%. \tag{5}$$

This splits the sample into 51 stressed years and 103 calm ones. Classification uses equity volatility over the full 1872–2025 history; including cash would shorten the regime record to the 44 years of T-bill data and leave too few transitions to estimate $P$.

*Stage 2 — transition counting.* With every year labelled, the one-step transition matrix is estimated by counting how often each move actually occurred. Let $n_{ij}$ be the number of years that started in regime $i$ and ended in regime $j$. The estimated probability of moving from $i$ to $j$ is then that count as a share of all moves out of $i$:

$$\hat{P}_{ij} \;=\; \frac{n_{ij}}{\sum_{j'} n_{ij'}}, \qquad \text{where} \qquad n_{ij} \;=\; \text{the number of years } t \text{ with } R_{t-1} = i \text{ and } R_t = j. \tag{6}$$

For example, of the 103 calm years that were followed by another year, 75 stayed calm, giving $\hat{P}_{11} = 75/103 = 0.728$.

Two alternatives were fitted and rejected. A two-component Gaussian mixture on annual *returns* simply bisects the return distribution near its mean and yields $P \approx \left[\begin{smallmatrix} .52 & .48 \\ .58 & .42 \end{smallmatrix}\right]$ against a stationary distribution of (.55, .45) — that is, essentially no persistence, which would remove the very clustering that Section 1 argues is the point of the exercise. The second alternative, a Markov-switching model estimated by expectation–maximisation (EM), is strongly persistent but identifies the wrong feature: its "stressed" state has *lower* equity volatility, higher bond volatility, and durations of 11–15 years, i.e. it detects the pre/post-1970s structural break in bond volatility, contradicting the no-structural-breaks assumption above. The volatility threshold avoids both failures, and its labels have strong face validity — they select 1873, 1893, 1907, 1929–35, 1937, 1974, 1987, 1998, 2001–02, 2008–09, 2011, 2020 and 2022. Results are stable for any threshold between roughly the 60th and 80th percentile.

**Fitted parameters.** Of the 154 years, 51 (33%) are classified stressed. The estimated transition matrix and its stationary distribution are

$$\hat{P} \;=\; \begin{pmatrix} 0.728 & 0.272 \\ 0.540 & 0.460 \end{pmatrix}, \qquad \hat{\pi} \;=\; (0.665,\; 0.335), \tag{7}$$

where row/column 1 is calm and 2 is stressed. Calm periods are persistent (expected duration $1/(1 - \hat{P}_{11}) = 3.7$ years); stressed periods are shorter (1.9 years). Table 1 reports the regime-conditional means, volatilities and correlations.

Because the cash series starts only in 1981, the equity and bond moments are estimated on the full 1872–2025 sample while the cash mean, volatility, and its correlations with equities and bonds are estimated on the 1982–2025 overlap; correlations rather than covariances are carried across, so that the long-sample equity and bond volatilities are preserved.

Table 1: Fitted regime-conditional parameters for real annual returns. Equity and bond moments are estimated on 1872–2025 (103 calm / 51 stressed years); cash moments and all correlations involving cash on 1982–2025 (32 calm / 12 stressed years).

| | **Calm regime** | | **Stressed regime** | |
|---|---|---|---|---|
| | Mean | Std. dev. | Mean | Std. dev. |
| S&P 500 | +12.64% | 15.38% | +0.56% | 20.06% |
| 10-year bonds | +2.54% | 7.67% | +3.26% | 10.52% |
| 3-month T-bills | +0.99% | 2.53% | +0.56% | 3.02% |
| *Correlations* | | | | |
| Equity–bond | +0.233 | | +0.119 | |
| Equity–cash | +0.120 | | +0.189 | |
| Bond–cash | +0.536 | | +0.673 | |

The stressed regime has a lower mean and higher volatility than the calm regime, as expected. It does *not* have higher equity–bond correlation: that correlation *falls* from +0.23 to +0.12. This runs against the common claim that diversification fails in crises, and the reason is specific to US Treasuries — investors buy them when equities sell off ("flight to quality"), so they have historically diversified better precisely when a retiree needs it most. This works in favour of the bond-holding scenarios in Section 4.

### 2.5 Decision Variables

- **Allocation policy**: a fixed-weight 60/40 equity/bond mix, versus an age-based glide path in which the equity share declines linearly over the horizon. The glide path is defined to start at the same 60/40 point so that any difference in results is attributable to the gliding itself:

$$w_t^{\text{eq}} = 0.60 - 0.30\,\frac{t-1}{T-1}, \qquad w_t^{\text{csh}} = 0.10\,\frac{t-1}{T-1}, \qquad w_t^{\text{bd}} = 1 - w_t^{\text{eq}} - w_t^{\text{csh}}, \tag{8}$$

  so equity falls 60% → 30%, cash rises 0% → 10%, and bonds absorb the remainder (40% → 60%).

- **Withdrawal rule**: a fixed real dollar amount, or a fixed percentage of the current balance,

$$W_t = \phi B_0 \quad \text{(fixed real \$)}, \qquad \text{or} \qquad W_t = \phi B_{t-1} \quad \text{(fixed \% of balance)}, \tag{9}$$

  with $\phi$ the withdrawal rate. A dynamic "guardrails" rule was considered and placed out of scope for this project; it is noted as an extension in Section 5.

- **Rebalancing frequency**: annual rebalancing back to target weights, versus no rebalancing (buy-and-hold), in which the weights drift with realized performance.

### 2.6 Performance Measures

- Probability of portfolio depletion ("ruin") before year 30, $\Pr(\tau \le T)$.
- Median and 10th-percentile terminal real wealth.
- Sustainable withdrawal rate for a 90% target success rate.
- Conditional Value at Risk (CVaR) of terminal wealth — the mean terminal wealth in the worst $\alpha = 5\%$ of replications.
- Volatility of annual real withdrawals (for percentage-of-balance rules), reported as a coefficient of variation.

**A common success criterion.** Ruin alone cannot rank the two withdrawal families against each other. A percentage-of-balance rule takes a fraction of what remains, so the balance approaches zero without reaching it and $\Pr(\tau \le T) \approx 0$ by construction — which flatters those rules without saying anything about the standard of living they deliver. We therefore define a path as *successful* when

$$\tau > T \quad \text{and} \quad \min_{1 \le t \le T} W_t \;\ge\; \kappa\, W_1, \qquad \kappa = 0.5, \tag{10}$$

that is, the portfolio never depletes *and* real income never falls below half of its first-year level. Under a fixed real-dollar rule the withdrawal is constant until the money runs out, so the second condition can never be the one that fails: Equation (10) reduces exactly to "not ruined", and the two families remain directly comparable. The sustainable withdrawal rate is the $\phi$ at which the success rate equals 90%.

# 3 Simulation Model

The simulation is implemented in Python (see Appendix A for the model code). Figure 2 summarizes the logic for a single scenario; this is repeated for every scenario in the experimental design (Section 4).

Figure 2: Monte Carlo simulation of one retirement scenario.

1. **for** $n = 1$ to $N = 10{,}000$ **do**
2.     Initialize regime $R_0$ (drawn from the stationary distribution $\pi$)
3.     Initialize balance $B_0 \leftarrow$ initial portfolio value
4.     **for** $t = 1$ to $T = 30$ **do**
5.         Transition regime: draw $R_t$ from row $R_{t-1}$ of $P$
6.         Draw asset returns $\mathbf{r}_t \sim \mathcal{N}(\boldsymbol{\mu}_{R_t}, \Sigma_{R_t})$ (equity, bond, cash)
7.         Compute portfolio return $\rho_t = \mathbf{w}_t^\top \mathbf{r}_t$ from current weights
8.         Determine withdrawal $W_t$ from the scenario's withdrawal rule
9.         Update balance: $B_t \leftarrow (B_{t-1} - W_t)(1 + \rho_t)$
10.         **if** rebalancing is active **then** reset weights to target allocation
11.         **if** $B_t \leq 0$ **then** record ruin at year $t$ and break
12.     **end for**
13.     Record terminal balance $B_T$ (or 0 if ruined) and the ruin indicator
14. **end for**
15. Aggregate across the $N$ replications into the performance measures of Section 2.6

**Implementation note.** The code tracks the three asset holdings separately rather than a single balance and a portfolio return. For rebalanced scenarios the two formulations are identical, since resetting to target weights each year makes the portfolio return exactly the weighted average in step 7. For the buy-and-hold scenario they are not: weights drift with realized performance, so the holdings must be carried individually. Each year's withdrawal is therefore funded by selling the same fraction of every holding, which leaves the portfolio's weights unchanged by the withdrawal itself.

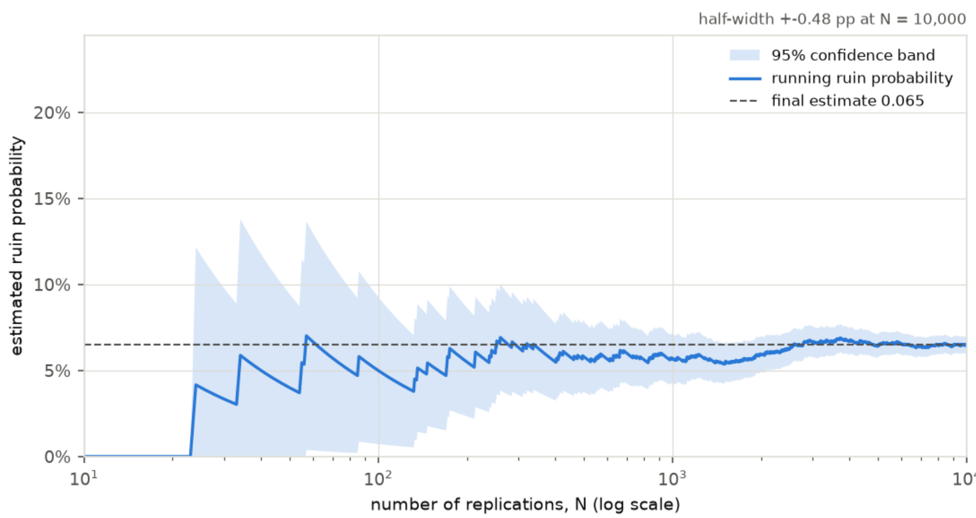
**Variance reduction.** All five scenarios are evaluated on the *same* 10,000 simulated market paths (common random numbers). Scenario comparisons are therefore paired, and differences between scenarios reflect the policies rather than differences in sampling noise.

**Choice of $N$.** Following the run-length-control ideas from Lecture 3, the number of replications is chosen so that the half-width of a 95% confidence interval on the ruin probability is acceptably small. For a proportion the half-width is

$$h \;=\; z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}. \tag{11}$$

For the baseline scenario, $\hat{p} = 0.0648$ at $N = 10{,}000$ gives $h = 0.0048$, i.e. $\pm 0.48$ percentage points — inside the $\pm 0.5$ pp target, though not by a wide margin. Inverting Equation (11) for $h = 0.005$ gives $N = 9{,}312$, so $N = 10{,}000$ is adequate and no increase is required. Note that precision improves only with $\sqrt{N}$: halving the interval would require $N \approx 40{,}000$. Figure 3 shows the estimate stabilizing as $N$ grows.

Figure 3: Convergence of the estimated ruin probability (with 95% confidence band) as the number of Monte Carlo replications increases, for the baseline scenario.




## 4 Experimental Analysis

We compare a small set of deliberately chosen scenarios rather than a full factorial grid, to keep the analysis focused. Table 2 lists the design. Both withdrawal rates are set to 4%, the classic benchmark from Bengen (1994): the fixed-dollar rules withdraw 4% of the *initial* balance each year in real terms, and the percentage rules withdraw 4% of the *current* balance, so that the first-year withdrawal is identical under both and the comparison is like-for-like.

Table 2: Scenario design for the experimental analysis.

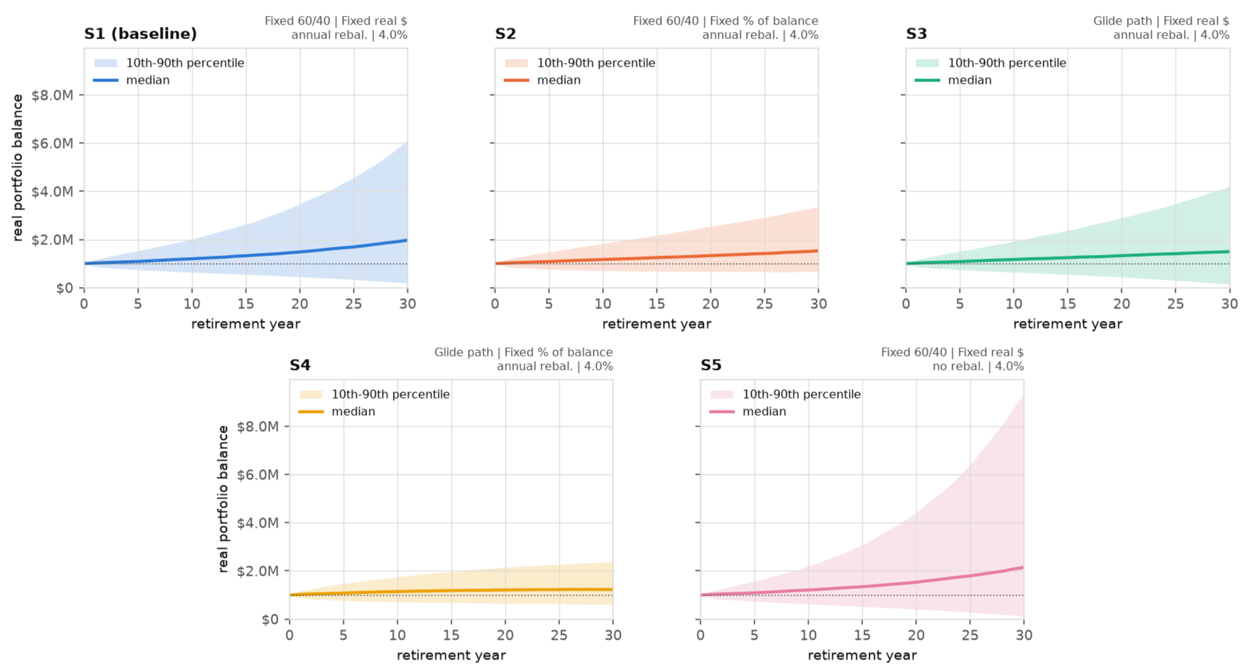
| Scenario | Allocation | Withdrawal Rule | Rebalancing |
|---|---|---|---|
| S1 (baseline) | Fixed 60/40 equity/bond | Fixed real \$, 4.0% | Annual |
| S2 | Fixed 60/40 equity/bond | Fixed % of balance, 4.0% | Annual |
| S3 | Glide path (equity declines w/ age) | Fixed real \$, 4.0% | Annual |
| S4 | Glide path (equity declines w/ age) | Fixed % of balance, 4.0% | Annual |
| S5 | Fixed 60/40 equity/bond | Fixed real \$, 4.0% | None |

These five were chosen so that each comparison isolates a single decision. S1 is the baseline — the classic 60/40 portfolio with a 4% real withdrawal, rebalanced annually. S1 against S5 varies *only* rebalancing, and so measures what rebalancing is worth. S1 against S3 varies only the allocation policy, isolating the glide path. S1 against S2, and S3 against S4, vary only the withdrawal rule. A full factorial design would add little: the remaining cells are combinations of effects that these four pairwise contrasts already identify, and each extra scenario makes the results table harder to read.

## 4.1 Distribution of Outcomes

Figure 4 shows the characteristic asymmetry of the problem: median wealth rises in every scenario, but the 10th-percentile band is flat or declining, so the typical retiree ends richer than they started while the unlucky decile is steadily eroded. The buy-and-hold scenario S5 has by far the widest band — unchecked equity drift produces both the best upside and the weakest downside. The percentage-of-balance scenarios (S2, S4) have visibly narrower bands, which is the same effect seen from the other side: spending adjusts downward automatically, which protects the portfolio at the cost of the retiree's income.

Figure 4: Fan chart of simulated portfolio balance over the 30-year horizon (median and 10th/90th percentile bands) for each scenario in Table 2. All five panels share both axes, so the bands can be compared directly across scenarios.




## 4.2 Ruin Probability and Sustainable Withdrawal Rates

Table 3 collects the headline measures. Reading across it, no scenario wins on every column. S5 has the highest median terminal wealth but also the highest ruin probability and the weakest 10th percentile; S2 and S4 never deplete but end with the least money in the median case. The baseline S1 is the only scenario that is close to the front on both counts, which is why the recommendations in Section 5 start from it.

Table 3: Summary performance measures by scenario. $N = 10{,}000$ replications; ruin probabilities carry a 95% CI half-width of at most $\pm 0.52$ pp.

| Scenario | Ruin Prob. | Median Terminal Wealth | 10th Pct. Terminal | CVaR (5%) | Sustainable WD Rate |
|---|---|---|---|---|---|
| S1 (baseline) | 6.48% | $1,972,063 | $189,875 | $0 | 4.34% |
| S2 | 0.00% | $1,530,331 | $665,764 | $393,533 | 3.93% |
| S3 | 6.90% | $1,494,219 | $141,100 | $0 | 4.29% |
| S4 | 0.00% | $1,225,988 | $609,816 | $390,930 | 4.08% |
| S5 | 7.45% | $2,146,184 | $122,738 | $0 | 4.23% |

CVaR is $0 for the three fixed-dollar scenarios because each has a ruin probability above the 5% tail cut-off, so the worst 5% of replications are all depleted paths. This is itself informative: it says the bad tail of a fixed-dollar rule is not "a small balance" but "nothing at all".

Figure 5 plots the ruin column with its confidence intervals. The three fixed-dollar scenarios sit between 6.5% and 7.5%, and their confidence intervals overlap only narrowly, so we can quite confidntly order the ruin probabilities S1 < S3 < S5. The two percentage-of-balance scenarios sit at zero, but as Section 2.6 explains that is guaranteed by the rule rather than earned by it, which is why the figure labels them explicitly.

Figure 5: Bar chart comparing the probability of portfolio depletion before year 30 across scenarios. The percentage-of-balance scenarios cannot deplete by construction.

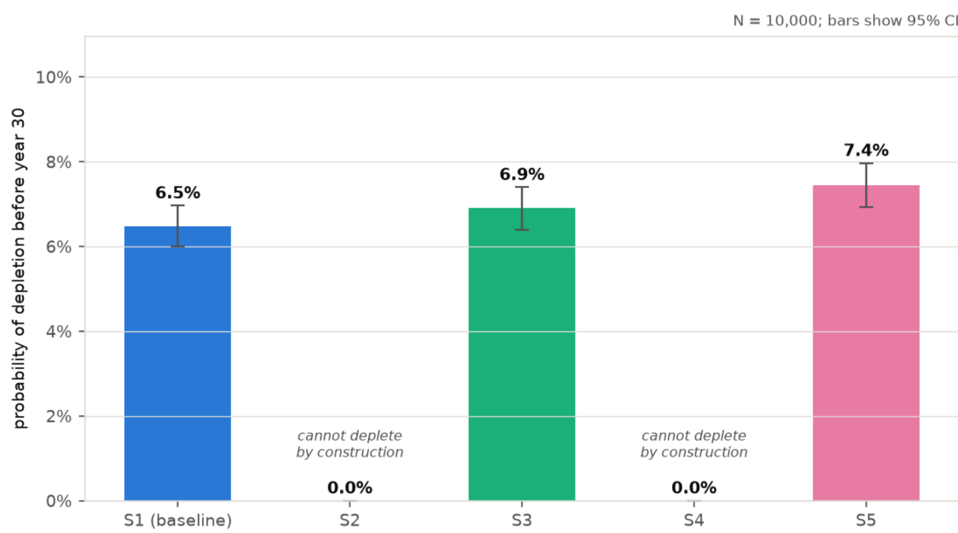
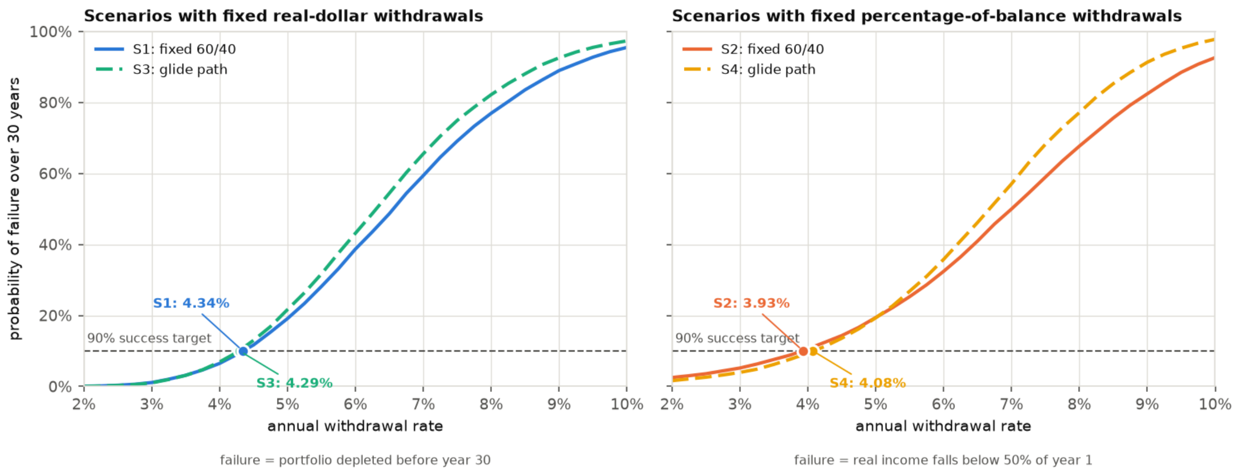



Figure 6 sweeps the withdrawal rate to find the sustainable rate directly. Each curve is one of the scenarios of Table 2 with its withdrawal rate varied, drawn in that scenario's colour, so the marked point where a curve crosses the target line is that scenario's sustainable rate. Failure means depletion for the fixed real-dollar scenarios S1 and S3 (left panel) and the income-floor criterion of Equation (10) for the percentage-of-balance scenarios S2 and S4 (right panel). The two panels cover the four rebalanced scenarios; The sustainable rate for the buy-and-hold scenario S5 is reported in Table 3. Both curves are steep through this range: for the baseline allocation, moving one percentage point above the sustainable rate — from 4.34% to 5.34% — lifts the failure probability from 10% to 25%. That steepness is the practical message of the figure. The cost of over-withdrawing is not gradual, so a retiree who sets the rate a little too high is not slightly worse off but substantially so. The allocation policy shifts the curve far less than the withdrawal rate does, which is consistent with the small gap between S1 and S3 in Table 3.

Figure 6: Failure probability as a function of the withdrawal rate, for each of the four rebalanced scenarios of Table 2, used to back out the withdrawal rate consistent with a 90% success rate. Line style separates the two allocation policies.




## 4.3 Discussion

**The 4% rule roughly holds up, and the model validates against the literature.** The baseline fails in 6.5% of retirements and supports a sustainable withdrawal rate of 4.34% at a 90% success target. Both figures sit squarely in the range reported by Bengen (1994) and the Trinity study (Cooley et al., 1998), which were built on historical bootstrapping rather than a regime-switching model. Arriving at the same answer by a different route is useful evidence that the result comes from the data rather than from our modelling choices.

**Rebalancing helps, but not by making more money.** S1 and S5 are identical except for rebalancing, so the difference between them is attributable to that alone. Rebalancing lowers the failure rate from 7.5% to 6.5%. Yet buy-and-hold has the *higher* median terminal wealth ($2.15M against $1.97M), because in the many paths where equities perform well, never trimming them back compounds faster. The difference is entirely in the left tail: S5's 10th-percentile outcome is $122,738 against S1's $189,875. Rebalancing is therefore best understood not as a return-enhancing device but as a mechanism that prevents the portfolio from drifting to maximum equity exposure just before a stressed regime arrives. For a retiree, who gets exactly one draw from this distribution, trading terminal wealth for tail protection seems like a worthy compromise.

**The glide path did not work, and the regime model explains why.** Shifting out of equities with age is standard advice, but here it slightly *raises* the failure rate (6.9% versus 6.5%) and clearly lowers median wealth ($1.49M versus $1.97M). The reason is likely attributable to the sequence-of-returns risk: the danger concentrates in the early years, when a stressed regime forces selling from a full-size portfolio. The linear glide path is still at its full 60% equity weight during exactly that window, and only reaches its defensive 30% weight around year 30, by which point most surviving portfolios are far enough ahead that the extra bond protection is redundant. It pays the cost of de-risking without buying much of the benefit. This suggests that if a glide path is used at all, it should be front-loaded — starting defensive and *increasing* equity exposure as the horizon shortens, the so-called rising-equity glide path — rather than the conventional declining one.

**Percentage-of-balance rules relocate risk rather than removing it.** S2 and S4 never deplete, but this is just arithmetic sincde 4% of a shrinking balance is always positive. The cost appears in the income stream, whose coefficient of variation is 24.3% for S2 and 19.4% for S4. This means that the retiree's real spending swings by roughly a quarter of its own average, and falls hardest exactly when markets are worst. Once the income floor of Equation (10) is imposed, S2's sustainable rate (3.93%) is *lower* than the baseline's (4.34%), which reverses the ranking the ruin column alone would suggest. Their strength is the opposite tail: S2 and S4 have far better 10th-percentile terminal wealth ($665,764 and $609,816) and are the only scenarios with non-zero CVaR. The honest summary is that they trade a small chance of catastrophe for a large chance of moderate disappointment. Which of those a retiree should prefer depends on their circumstances; neither rule is better on every measure.

**Does the regime structure change the conclusions?** A single-normal model fitted to the same data would use the pooled equity mean of 8.6% and volatility of 17.9%, and would draw each year independently. The two-regime model has a similar unconditional mean but generates *clustered* bad years: because the calm state persists with probability 0.73 and the stressed state with probability 0.46, runs of poor returns occur far more often than independent sampling would produce. Since ruin is driven by consecutive early losses rather than by any single bad year, the regime model produces a materially fatter left tail at the same average return. It also explains why the ranking of policies depends on *when* risk is taken — the mechanism behind the glide-path result above, and one a single-normal model could not have revealed.

# 5 Conclusions and Recommendations

For a retiree with a 30-year horizon, the analysis supports a fixed 60/40 portfolio, rebalanced annually, with an initial real withdrawal rate of about 4.3%. Specifically, we recommend the retiree to:

1. **Rebalance.** It is the cheapest improvement available: a one-percentage-point reduction in failure probability and a 55% improvement in the 10th-percentile outcome, at no financial cost.
2. **Do not adopt a conventional declining glide path.** On this model it costs both expected wealth and safety, because it de-risks after the window of maximum vulnerability has closed.
3. **Choose the withdrawal rate and rule according to which risk is less tolerable.** Assuming 60/40 and rebalancing, fixed real dollar withdrawals protect the standard of living and accept a 6.5% chance of depletion. A percentage-of-balance rule eliminates depletion but delivers income that varies by roughly a quarter of its average. Neither dominates; retirees with substantial non-portfolio income (e.g. Social Security, a pension) can absorb the variability and might prefer the percentage rule, while those whose portfolio is a way of funding essential spending should prefer the fixed rule.

4. **Treat 4% as more of an upper bound.** It clears the 90% success bar only before taxes and fees, and only under the assumption that the next 30 years resemble the last 154.

**Limitations.** The model assumes no taxes, fees, or transaction costs, which makes every withdrawal rate quoted here optimistic. Regime and distributional parameters are estimated once from history and held fixed for the full 30-year horizon; using the most successful equity market in history to forecast the future carries a well-known survivorship bias. The cash/T-bill leg is estimated from a shorter (1981–present) sample than equities and bonds (1871–present) — only 12 of its 44 years are stressed — so its regime-conditional parameters are the least reliable inputs, though cash is also the smallest holding in every scenario. Returns within a regime are assumed multivariate normal, which remains a simplification even though the regime mixture produces fat tails overall. Finally, the two-stage fitting procedure treats the regime labels as certain when they are themselves estimates, which most likely understates the true persistence of the regimes.

**Possible extensions.** A third (crisis) regime with its own dynamics, to separate brief violent episodes such as 1929–33 and 2008 from ordinary turbulence; a rising-equity glide path, which the discussion above suggests would outperform the conventional declining one; a full guardrails withdrawal rule with both upward and downward adjustments, which would sit between the two rule families tested here; incorporating longevity risk instead of a fixed 30-year horizon; adding international equity or real assets as further diversifiers; and taxes and fees, which would lower sustainable withdrawal rates in practice.

## References


1. Shiller, R. J. *Online Data — U.S. Stock Markets 1871–Present and CAPE Ratio.* `http://www.econ.yale.edu/~shiller/data.htm`
2. Federal Reserve Bank of St. Louis. *3-Month Treasury Bill Secondary Market Rate (DGS3MO).* FRED. `https://fred.stlouisfed.org/series/DGS3MO`
3. Bengen, W. P. (1994). Determining Withdrawal Rates Using Historical Data. *Journal of Financial Planning*, 7(4), 171–180.
4. Cooley, P. L., Hubbard, C. M., & Walz, D. T. (1998). Retirement Savings: Choosing a Withdrawal Rate That Is Sustainable. *AAII Journal*, 20(2), 16–21.
5. Course notes, *Stochastic Modeling and Simulation* (SHBI-GB.7301.B10), Lectures 3, 4, 7–8.

# A Code

The simulation model is submitted as `simulation_s.py`, together with `regime_params.json`, the fitted parameter file it reads at start-up. Reported below is the code that defines the model: the regime fit that produces its inputs (Appendix A.1) and the simulation itself (Appendices A.2–A.5). Data preparation, figure generation, and console reporting are not reported — they are mechanical and add nothing to the specification in Sections 2 and 3. Random seeds are fixed, so the reported numbers reproduce exactly.

## A.1 Fitting the regime model

Classification of each year into calm or stressed per Equation (5), estimation of the transition matrix and its stationary distribution, and construction of the regime-conditional mean vectors and covariance matrices reported in Table 1. The input `annual` is a table of real annual returns for the three assets; `monthly` holds the underlying monthly index levels.

```
CALM, STRESSED = 0, 1
STRESS_QUANTILE = 2 / 3

# Realised volatility: how choppy each year was, not whether it ended up or down.
monthly_equity_return = monthly["equity"].pct_change()
realized_vol = (monthly_equity_return
                .groupby(monthly_equity_return.index.year)
                .std(ddof=1)
                .reindex(annual.index))

vol_threshold = realized_vol.loc[LONG_YEARS].quantile(STRESS_QUANTILE)
regimes = (realized_vol.loc[LONG_YEARS] > vol_threshold).astype(int)


def transition_matrix(state_series, n_states=2):
    """Count year-to-year transitions and row-normalise into P."""
    counts = np.zeros((n_states, n_states))
    prev, nxt = state_series.to_numpy()[:-1], state_series.to_numpy()[1:]
    for i, j in zip(prev, nxt):
        counts[i, j] += 1
    return counts / counts.sum(axis=1, keepdims=True)


def stationary_distribution(P):
    """Left eigenvector of P for eigenvalue 1, normalised to sum to 1."""
    vals, vecs = np.linalg.eig(P.T)
    pi = np.real(vecs[:, np.argmin(np.abs(vals - 1))])
    return pi / pi.sum()


def regime_parameters(annual, regimes, long_years, overlap_years, state):
    """Mean vector and covariance matrix for one regime.

    Equity and bond moments come from the full 1872-2025 history. The cash
    moments, and the correlations involving cash, come from the shorter window
    where T-bill data exists; carrying correlations rather than covariances
    across preserves the long-sample equity and bond volatilities.
    """
    long_rows = annual.loc[long_years[regimes.loc[long_years] == state],
                           ["equity", "bond"]]
    over_rows = annual.loc[overlap_years[regimes.loc[overlap_years] == state], ASSETS]

    mu = np.empty(3)
    mu[:2] = long_rows.mean().to_numpy()
    mu[2] = over_rows["cash"].mean()
```

```python
    sd = np.empty(3)
    sd[:2] = long_rows.std(ddof=1).to_numpy()
    sd[2] = over_rows["cash"].std(ddof=1)

    corr = np.eye(3)
    corr[0, 1] = corr[1, 0] = long_rows.corr().to_numpy()[0, 1]
    short_corr = over_rows.corr().to_numpy()
    corr[0, 2] = corr[2, 0] = short_corr[0, 2]
    corr[1, 2] = corr[2, 1] = short_corr[1, 2]

    return mu, corr * np.outer(sd, sd)
```

## A.2 Scenario definitions and allocation policies

The five scenarios of Table 2, and the glide path of Equation (8).

```python
N_REPS = 10_000
T_YEARS = 30
B0 = 1_000_000.0
FIXED_DOLLAR_RATE = 0.04   # fraction of the INITIAL balance, constant in real terms
FIXED_PCT_RATE = 0.04      # fraction of the CURRENT balance, re-struck each year
INCOME_FLOOR_FRAC = 0.50   # income floor used by the success criterion


def target_weights(policy, T=T_YEARS):
    """(T, 3) matrix of target weights, one row per retirement year."""
    if policy == "fixed_60_40":
        return np.tile([0.60, 0.40, 0.00], (T, 1))
    if policy == "glide_path":
        # Starts at the same 60/40 as the baseline, then moves into bonds and cash.
        frac = np.arange(T) / (T - 1)
        equity = 0.60 - 0.30 * frac
        cash = 0.10 * frac
        return np.column_stack([equity, 1.0 - equity - cash, cash])
    raise ValueError(f"unknown allocation policy: {policy}")


SCENARIOS = {
    "S1": dict(allocation="fixed_60_40", rule="fixed_real",
               rate=FIXED_DOLLAR_RATE, rebalance=True),
    "S2": dict(allocation="fixed_60_40", rule="fixed_pct",
               rate=FIXED_PCT_RATE, rebalance=True),
    "S3": dict(allocation="glide_path", rule="fixed_real",
               rate=FIXED_DOLLAR_RATE, rebalance=True),
    "S4": dict(allocation="glide_path", rule="fixed_pct",
               rate=FIXED_PCT_RATE, rebalance=True),
    "S5": dict(allocation="fixed_60_40", rule="fixed_real",
               rate=FIXED_DOLLAR_RATE, rebalance=False),
}
```

## A.3 The market generator

Steps 2, 5 and 6 of Figure 2: regime paths from the Markov chain of Equation (3), and correlated regime-conditional returns from Equation (4).

```python
def simulate_market(n_reps, T, rng):
    """Draw regime paths and the returns conditional on them.

    Generated once and reused across all five scenarios (common random numbers),
    so scenario comparisons are paired rather than confounded by sampling noise.
    """
    n_states = len(PI)
    state = rng.choice(n_states, size=n_reps, p=PI)   # start from stationary
```

```
    regimes = np.empty((n_reps, T), dtype=np.int8)

    for t in range(T):
        u = rng.random(n_reps)
        state = (u[:, None] > np.cumsum(P[state], axis=1)).sum(axis=1)
        regimes[:, t] = state

    # r = mu_k + L_k z, with L_k the Cholesky factor of that regime's covariance.
    z = rng.standard_normal((n_reps, T, len(ASSETS)))
    returns = np.empty_like(z)
    for k in range(n_states):
        mask = regimes == k
        returns[mask] = MU[k] + z[mask] @ np.linalg.cholesky(SIGMA[k]).T
    return regimes, returns
```

## A.4 The retirement ledger

Steps 3 and 7–13 of Figure 2, vectorised across replications.

```
def run_scenario(returns, allocation, rule, rate, rebalance, B0=B0):
    """Run one scenario across every replication at once.

    The three holdings are tracked separately rather than as a single balance.
    Under annual rebalancing this is equivalent to applying a weighted-average
    portfolio return, but for buy-and-hold it is not: the weights drift with
    performance, which is one of the effects the experiment is measuring.
    """
    n_reps, T, _ = returns.shape
    weights = target_weights(allocation, T)

    holdings = np.tile(B0 * weights[0], (n_reps, 1))
    balances = np.zeros((n_reps, T + 1))
    balances[:, 0] = B0
    withdrawals = np.zeros((n_reps, T))
    ruin_year = np.zeros(n_reps, dtype=int)
    alive = np.ones(n_reps, dtype=bool)

    for t in range(T):
        balance = holdings.sum(axis=1)

        if rule == "fixed_real":
            want = np.full(n_reps, rate * B0)
        elif rule == "fixed_pct":
            want = rate * balance
        else:
            raise ValueError(f"unknown withdrawal rule: {rule}")
        want = np.where(alive, want, 0.0)

        # Ruin is the first year the balance cannot fund the full withdrawal.
        newly_ruined = alive & (balance <= want)
        taken = np.minimum(want, balance)

        share = np.divide(taken, balance, out=np.zeros(n_reps), where=balance > 0)
        holdings *= (1.0 - share)[:, None]
        holdings *= np.maximum(1.0 + returns[:, t, :], 0.0)

        if rebalance:
            holdings = holdings.sum(axis=1)[:, None] * weights[min(t + 1, T - 1)]

        ruin_year[newly_ruined] = t + 1
        alive &= ~newly_ruined
        holdings[~alive] = 0.0

        withdrawals[:, t] = taken
        balances[:, t + 1] = holdings.sum(axis=1)

```

```
    return dict(balances=balances, withdrawals=withdrawals,
                terminal=balances[:, -1], ruin_year=ruin_year,
                ruined=ruin_year > 0)
```

## A.5 Performance measures

The measures of Section 2.6, including the success criterion of Equation (10) and the bisection used to locate the sustainable withdrawal rate plotted in Figure 6.

```
from scipy.stats import norm


def success_mask(res):
    """Never depleted, and real income never fell below half of year one."""
    first = res["withdrawals"][:, 0]
    worst = res["withdrawals"].min(axis=1)
    return (~res["ruined"]) & (worst >= INCOME_FLOOR_FRAC * first)


def ruin_ci(ruined, confidence=0.95):
    """Normal-approximation CI half-width for a proportion."""
    n = len(ruined)
    p = ruined.mean()
    z = norm.ppf(0.5 + confidence / 2)
    return p, z * np.sqrt(p * (1 - p) / n)


def cvar(terminal, alpha=0.05):
    """Expected terminal wealth in the worst alpha tail."""
    cutoff = max(1, int(np.floor(alpha * len(terminal))))
    return np.sort(terminal)[:cutoff].mean()


def failure_probability(allocation, rule, rate, rebalance):
    res = run_scenario(RETURNS, allocation, rule, rate, rebalance)
    return 1.0 - success_mask(res).mean()


def sustainable_rate(allocation, rule, rebalance, target=0.90,
                     lo=0.005, hi=0.20, tol=1e-4):
    """Bisect for the withdrawal rate whose success rate equals the target.

    Failure probability is monotone increasing in the rate, and every evaluation
    reuses the same market paths, so the root is well defined.
    """
    target_failure = 1.0 - target
    while hi - lo > tol:
        mid = 0.5 * (lo + hi)
        if failure_probability(allocation, rule, mid, rebalance) < target_failure:
            lo = mid
        else:
            hi = mid
    return 0.5 * (lo + hi)
```