Stockholms universitet

# Forecasting and Backtesting Value-at-Risk

Sherin Shabo

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2025:2
Matematisk statistik
Februari 2025

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2025:2**
http://www.math.su.se

# Forecasting and Backtesting Value-at-Risk

Sherin Shabo*

February 2025

**Abstract**

The main focus of this thesis lies on the forecasting and backtesting of Value-at-Risk (VaR), a widely used metric for assessing financial risk. Using ten years of daily returns for the S&P 500 index, various methods are applied to estimate VaR at 95% and 99% confidence levels. These include traditional approaches like Historical Simulation (HS) and Variance-Covariance (VC) methods, as well as advanced methods incorporating time-varying volatility, such as Exponentially Weighted Moving Average (EWMA) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH). The study reveal that conditional methods, particularly EWMA-based approaches, consistently outperform traditional models, adapting effectively to changing market conditions and delivering more accurate risk estimates. Backtesting results using the Kupiec and Christoffersen tests demonstrate the reliability of EWMA methods in both stable and volatile periods, especially during the heightened market turbulence of 2020. In contrast, HS-GARCH methods exhibit unusual behavior, likely due to inherent challenges in combining GARCH volatility modeling with the historical simulation approach, highlighting the need for further investigation.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: sherin.adam@gmail.com. Supervisor: Johannes Heiny, Xuechun Hu.

## Acknowledgments

This work constitutes a bachelor's thesis of 15 ECTS in Mathematical Statistics at the Department of Mathematics, Stockholm University.

I would like to express my heartfelt gratitude to my supervisors, Johannes Heiny and Xuechun Hu, for their invaluable support and guidance throughout this journey. I am especially grateful to Johannes for recommending an excellent book that served as the foundation for the theoretical part of this thesis. His constructive feedback, insightful suggestions, and help in identifying errors have been instrumental in shaping the structure and quality of this work.

I am profoundly thankful to my husband and best friend, Farhan Shabo, and my four wonderful children, Adam, Danyar, Ramatia, and Lydia, for their unwavering support and patience. Their belief in me, constant encouragement, and understanding of the challenges I faced were indispensable. For having them in my life, I am forever grateful.

My deep appreciation also goes to Esma Özdolap and Yasmin Abdullah for their kind support, and to my brother, Mamand Abdullah, for his valuable suggestions, thorough text reviews, and steadfast encouragement.

Finally, I wish to acknowledge the assistance of ChatGPT, which proved invaluable for grammar checking, providing code suggestions, and offering LaTeX support during this process.

# Contents

# 1 Introduction

Risk in finance refers to the possibility of experiencing a loss or an unfavorable outcome due to uncertainties in financial markets or other factors affecting the value of investments. It is an inherent aspect of financial markets, as prices of assets, such as stocks, bonds, or commodities, fluctuate due to various factors including economic events, geopolitical tensions, and market sentiment. Financial risk can be classified into different types, including market risk (the risk of changes in asset prices), credit risk (the risk that a borrower will default on a loan), liquidity risk (the risk of not being able to sell an asset without significant loss), and operational risk (the risk of failure in internal processes or systems)(see McNeil et al. [1], p.3).

To measure and manage risk as mentioned by Jorion ([2], p.248), financial professionals use various metrics and models such as standard deviation, Value-at-Risk (VaR) and expected shortfall (ES). Standard deviation measures the volatility or variability of asset returns, while VaR estimates the potential loss in a portfolio over a specific time period and confidence level. Expected shortfall (ES) also known as the conditional value at risk (CVaR) is a related concept to VaR and it refers to the expected value of the loss conditional on exceeding VaR. By accurately measuring risk, investors, banks, and firms can make informed decisions to mitigate potential losses and optimize their portfolios.

VaR is commonly used by financial institutions and investors to estimate potential losses in their portfolios under normal market conditions. It provides a single value that represents the overall risk of a portfolio ([3], p.255). JP Morgan introduced it in the late 1980s as part of its risk management system, developing a tool called RiskMetrics, which later became an independent company. Although JP Morgan did not create VaR, it helped make the concept more widely known and used in finance ([4], p. 160).

In this thesis, we focus on evaluating Value-at-Risk as a tool for managing financial risk. We apply various methods to forecast VaR for the S&P 500 index over a 10-year period and assess their performance using backtesting techniques. The analysis involve both traditional and advanced models, such as Historical Simulation, Variance-Covariance, EWMA, and GARCH, with a particular focus on how well these methods adapt to changing market conditions.

# 2 Theoretical Framework

## 2.1 General Definitions

In this section, we provide formal definitions and a consistent notation for the loss distribution and Value-at-Risk, using the quantile function as a tool. If not explicitly stated otherwise, the references for this section are Chapters 2.1 and 2.2 in [1] and Chapter 12.5 in [3].

### 2.1.1 Quantile Function

The quantile function provides a way to connect cumulative probabilities with the values of a random variable in a distribution. It acts as the generalized inverse of the cumulative distribution function (CDF)[5].

**Definition 2.1** (**Quantile function**). Let $F : \mathbb{R} \to [0, 1]$ be a cumulative distribution function. For a given $\alpha \in (0, 1)$, the $\alpha$-quantile of $F$ is given by

$$q_\alpha(F) := \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}.$$

In practice, for a random variable $X$ with distribution function $F$, we often use the shorthand notation $q_\alpha(X) := q_\alpha(F)$, which allows us to refer to the quantile function of $X$ in terms of the distribution function $F$. If $F$ is continuous and strictly increasing, we can simplify this expression to $q_\alpha(F) = F^{-1}(\alpha)$, where $F^{-1}$ is the inverse of $F$.
Figure 1 below illustrates this concept using the normal distribution, showing the relationship between the CDF and its inverse.

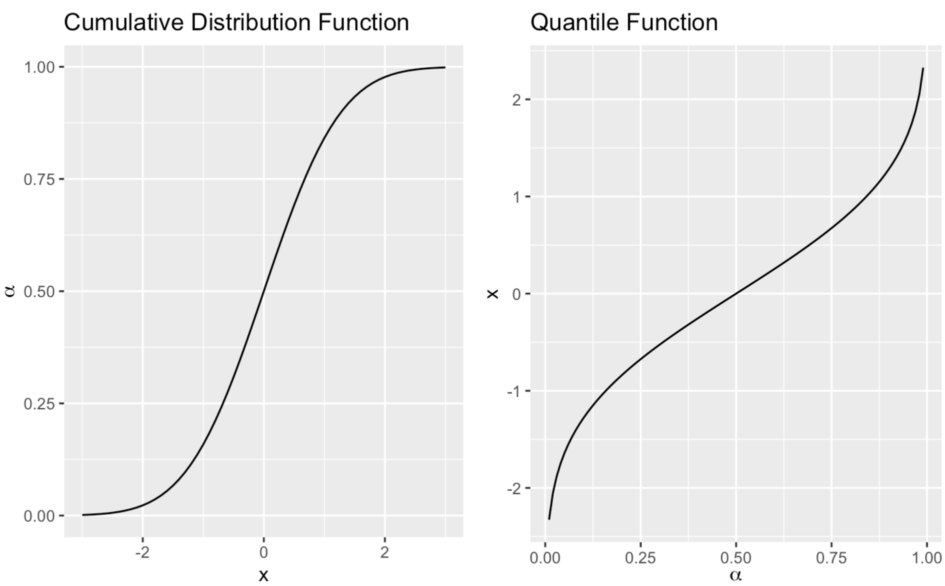



Figure 1: The cumulative distribution function $F$ (on the left) gives the probability $\alpha = F(x)$ for a given value $x$. The quantile function (on the right) does the opposite: it finds the value $x$ that solves $\alpha = F(x)$ for a given probability $\alpha$.

For example, if $\alpha = 0.5$, the quantile function identifies the median of the distribution. Similarly, the 25th percentile corresponds to $q_{0.25}(F)$, and the 75th percentile corresponds to $q_{0.75}(F)$. The quantile function is particularly helpful in determining such thresholds for a distribution [6].

### 2.1.2 Loss Distribution

The *loss distribution* describes the probability of different levels of loss that a portfolio may experience over a given time period.

**Definition 2.2** (**Loss Distribution**)**.** Let $\Delta$ be a fixed time horizon , such as 1 or 10 days and let $V_t$ represent the value of a portfolio at time $t\Delta$. The portfolio loss between times $t\Delta$ and $(t+1)\Delta$ is given by

$$L_{t+1} := -(V_{t+1} - V_t). \tag{1}$$

By this convention, losses will be positive numbers and profits negative. This loss is *random* from the perspective of time $t\Delta$ because the future value of the portfolio is unknown.
There are two types of loss distribution:

- The *unconditional loss distribution* describes the loss without considering any specific information about the state of the world at time $t\Delta$.
- The *conditional loss distribution* accounts for the information available at time $t\Delta$, i.e., the loss distribution given what is known at that time.

In practice, risk managers often focus on the upper tail of the loss distribution, which represents the probability of large losses. This focus on large losses is critical to managing risk, especially when analyzing worst-case scenarios.

### 2.1.3 Value-at-Risk

Consider a portfolio of risky assets over a fixed time horizon $\Delta$. Let $F_L(l) = P(L \leq l)$ represent the distribution function of losses over this time. The goal is to define a risk measure based on $F_L$ that can effectively assess the severity of holding this portfolio during $\Delta$.
An intuitive idea for a risk measure might be the *maximum possible loss*, represented as

$$\inf\{l \in \mathbb{R} : F_L(l) = 1\}.$$

However, this measure is problematic because in many financial models the support of $F_L$ is unbounded, making the maximum loss potentially infinite and lacking the probability context. Instead, Value-at-Risk (VaR) provides a more practical measure by considering the "maximum loss that is not exceeded with a given high probability." This probability is called the *confidence level*.

**Definition 2.3** (**Value-at-Risk**)**.** For a given confidence level $\alpha \in (0, 1)$, the VaR at a confidence level $\alpha$ is the smallest value $l$ such that the probability of loss $L$ exceeding $l$ is not greater than $1 - \alpha$. Formally,

$$\mathrm{VaR}_\alpha = \inf\{l \in \mathbb{R} : P(L > l) \leq 1 - \alpha\} = \inf\{l \in \mathbb{R} : F_L(l) \geq \alpha\}.$$

$\mathrm{VaR}_\alpha$ is a quantile of the loss distribution and to measure it, we need to define two quantitative parameters, the confidence interval and the time horizon. Typical confidence levels for $\alpha$ are 0.95 or 0.99 and for market risk, the time horizon $\Delta$ is often 1 or 10 days. The choice of these parameters is described in more detail in ([1], p.42) and ([2], p.251,252).

To better understand the concept of Value-at-Risk (VaR), let us consider a practical example. Suppose we are managing a portfolio and determine that $\mathrm{VaR}_{0.95} = \$1,000,000$ over a time horizon of 10 days. This means that, with 95% confidence, the portfolio is *not expected to lose more than $1,000,000* over the next 10 days. In other words, there is a 5% chance that the losses could exceed this amount within that 10-day period. The remaining 5% represents the *tail risk*, those extreme losses that are less frequent but still possible.

VaR gives an estimate of the worst loss we might expect, assuming normal market conditions, with a certain level of confidence. However, one of the key limitations of VaR is that it *does not tell us anything* about how bad the losses could be if they end up being worse than the estimate of VaR. In other words, it does not provide insight into how severe losses might be in the worst 5% cases, which means that it misses information about extreme risks.

Figure 2 illustrates a loss distribution with the Value-at-Risk (VaR) marked by a vertical dashed line. VaR represents the threshold loss not expected to be exceeded with a given confidence level (e.g. 95%). The area to the left of this line shows the probability that losses will remain below this threshold.

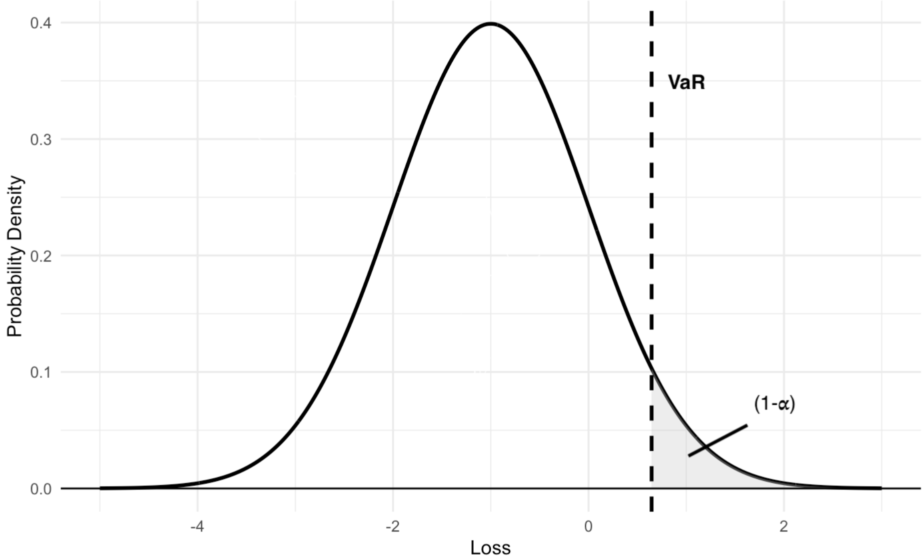



Figure 2: Calculating VaR from the distribution of the Loss in the Portfolio Value. The confidence interval is $\alpha$ and $\mathrm{VaR}_\alpha$ is marked with a dashed line. Gains are negative losses.

## 2.2 Standard Methods for Market Risks

In this section, we will present several commonly used methods in the financial industry for measuring market risk over a short time horizon, such as the variance covariance method, the historical simulation, and methods based on Monte Carlo simulation. Before introducing these three methods, we first need to model portfolio losses based on changes in risk factors. The references for this section are Chapters 2 and 3 in [1].

### 2.2.1 Modeling Portfolio Losses with Risk Factors

In risk management, the value of a portfolio at any time $t$, denoted by $V_t$, is often modeled as a function of time and a vector of *risk factors* $\boldsymbol{Z}_t = (Z_{t,1}, Z_{t,2}, \ldots, Z_{t,d})^\top$ . Typical risk factors include, for example, the logarithmic prices of financial assets, yields, or the logarithmic exchange rate. The portfolio value at time $t$ is represented as

$$V_t = f(t, \boldsymbol{Z}_t), \tag{2}$$

where $f$ is a measurable function that maps time and risk factors to the portfolio value. The risk factors are assumed to be observable, which means $\boldsymbol{Z}_t$ is known at any given time $t$. The choice of risk factors depends on the portfolio under consideration and the desired level of precision.

Using (1) and the mapping (2), the change in portfolio value (i.e., the *portfolio loss*) from time $t$ to $t+1$ is given by

$$L_{t+1} = -(f(t+1, \boldsymbol{Z}_t + \boldsymbol{X}_{t+1}) - f(t, \boldsymbol{Z}_t)), \tag{3}$$

where $\boldsymbol{X}_{t+1}$ represents the *change in risk factors* between time $t$ and $t+1$.
To model how risk-factor changes map into losses, a *loss operator* is introduced, which is defined as

$$l_{[t]}(\boldsymbol{x}) := -(f(t+1, \boldsymbol{Z}_t + \boldsymbol{x}) - f(t, \boldsymbol{Z}_t)), \quad \boldsymbol{x} \in \mathbb{R}^d. \tag{4}$$

Thus, we have that $L_{t+1} = l_{[t]}(\boldsymbol{X}_{t+1})$.
For small changes in risk factors, a *first-order Taylor expansion* is used to approximate portfolio loss (3). The linear approximation of the loss at time $t+1$ is given by

$$L^{\Delta}_{t+1} := -(f(t, \boldsymbol{Z}_t) + \sum_{i=1}^{d} f_{z_i}(t, \boldsymbol{Z}_t)\boldsymbol{X}_{t+1,i}). \tag{5}$$

Here, $f_{z_i}(t, \boldsymbol{Z}_t)$ represents the *partial derivatives* of $f$ with respect to each risk factor, indicating how sensitive the portfolio value is to changes in each risk factor. The term $\boldsymbol{X}_{t+1,i}$ represents the change in the $i$-th risk factor.
This *linearized loss* approximation simplifies the calculation of losses by considering the direct impact of small risk-factor changes, assuming that the relationship between the portfolio value and the risk factors is nearly linear for small changes.
To make the approximation computationally simpler, a *linearized version of the loss operator* is introduced

$$l^{\Delta}_{[t]}(\boldsymbol{x}) := -(f(t, \boldsymbol{Z}_t) + \sum_{i=1}^{d} f_{z_i}(t, \boldsymbol{Z}_t)x_i), \tag{6}$$

where $\boldsymbol{x} = (x_1, ....., x_d)^T$ represents changes in the risk factors. This approximation is especially useful when the changes in risk factors are small (i.e., when looking at risk over a short period), and when the portfolio value responds almost linearly to changes in the risk factors.

### 2.2.2 Variance-Covariance Method (VC)

The *Variance-Covariance method* is a commonly used technique for calculating Value-at-Risk in finance. This method simplifies the process by making several key assumptions. First, it assumes that the changes in the risk factors $\boldsymbol{X}_{t+1}$ follow a *multivariate normal distribution*, that is, $\boldsymbol{X}_{t+1} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. So the the changes in risk factors in this model, are represented by a random vector $\boldsymbol{X}_{t+1}$, which has a *mean vector* $\boldsymbol{\mu}$ and a *covariance matrix* $\boldsymbol{\Sigma}$.
Additionally, the method assumes that the linearized loss, represented in terms of the risk factors, is a reasonable approximation of the actual loss. This enables us to simplify the problem by examining the distribution of

$$L^{\Delta}_{t+1} = l^{\Delta}_{[t]}(\boldsymbol{X}_{t+1}),$$

where $l^{\Delta}_{[t]}$ is given in (6)
The linearized loss at time $t$ is expressed as

$$l^{\Delta}_{[t]}(\boldsymbol{x}) = -(c_t + \boldsymbol{b}'_t\boldsymbol{x}),$$

where $c_t$ is a constant, and $\boldsymbol{b}_t$ is a vector known at time $t$.
An important property of the *multivariate normal distribution* is that any *linear combination* of the components of the random vector $\boldsymbol{X}_{t+1}$ (such as $\boldsymbol{b}'_t\boldsymbol{X}_{t+1}$) will also follow a normal distribution. As a result, the loss $L^{\Delta}_{t+1}$ is normally distributed with the following mean and variance

$$L^{\Delta}_{t+1} = l^{\Delta}_{[t]}(\boldsymbol{X}_{t+1}) \sim \mathcal{N}(-c_t - \boldsymbol{b}_t'\boldsymbol{\mu}, \boldsymbol{b}_t'\boldsymbol{\Sigma}\boldsymbol{b}_t).$$

To calculate Value-at-Risk, we need to find the *quantile* of the loss distribution at a given confidence level $\alpha$. For this purpose we can use the following lemma.

**Lemma 2.1** (**VaR for normal loss distributions**)**.** Suppose the loss distribution $F_L$ is normal with mean $\mu$ and variance $\sigma^2$, and let $\alpha \in (0, 1)$. Then the Value-at-Risk at confidence level $\alpha$ is given by

$$\mathrm{VaR}_\alpha = \mu + \sigma\Phi^{-1}(\alpha),$$

where $\Phi$ denotes the standard normal distribution function, and $\Phi^{-1}(\alpha)$ is the $\alpha$-quantile of the standard normal distribution.
For a proof of this lemma, see ([1], p.39).

The Student's $t$-distribution is commonly used in risk modeling, particularly when dealing with heavy tails. The loss distribution is modeled as $L \sim t(\nu, \mu, \sigma^2)$, where $\nu$ represents the degrees of freedom, $\mu$ is the mean, and $\sigma^2$ is the variance of the distribution. The moments of the loss distribution are given by $E(L) = \mu$ and $\mathrm{var}(L) = \frac{\nu\sigma^2}{\nu-2}$, when $\nu > 2$.

**Lemma 2.2** (**VaR for the Student's t-Distribution**)**.** The Value-at-Risk for a loss modeled by a Student's $t$-distribution is calculated as

$$\mathrm{VaR}_\alpha = \mu + \sigma t_\nu^{-1}(\alpha),$$

where $t_\nu^{-1}(\alpha)$ is the inverse of the cumulative distribution function (CDF) of the $t$-distribution at the confidence level $\alpha$.

**Remark 2.1.** For the variance-covariance method to work, we require estimates for the mean vector $\mu$ and the covariance matrix $\Sigma$, which are derived from historical data. These estimates can be calculated using the standard method of moments estimators ([1], p. 65), specifically:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})^\top,$$

where $X_1, \ldots, X_n$ represents the observations, $n$ is the number of data points and $\hat{\mu}$ and $\hat{\Sigma}$ are the mean vector and the covariance matrix of the sample, respectively.

This approach amounts to an analysis of the unconditional loss distribution, which assumes stationarity in the data. Alternatively, more advanced methods can be employed to forecast the conditional moments of the distribution, such as exponentially weighted moving averages (EWMA) or GARCH models. These forecasting techniques are designed to account for time-varying characteristics in the data and are discussed in greater detail in Section 2.3.

### 2.2.3 Historical Simulation (HS)

The *Historical simulation method* is a non-parametric approach for estimating Value-at-Risk by using historical data on risk-factor changes. Instead of relying on a probabilistic model to estimate the distribution of $L = l_{[t]}(\boldsymbol{X}_{t+1})$, this method estimates the distribution directly using historical simulation.
Specifically, we gather historical data on the changes in risk factors. By applying the *loss operator* to this data, we calculate the simulated portfolio losses

$$\{\tilde{L}_s = l_{[t]}(\boldsymbol{X}_s) : s = t-n+1, \dots, t\},$$

which represent the portfolio loss that would occur if the risk-factor changes observed on day $s$ were to recur.
These simulated losses $\tilde{L}_s$ form an *empirical distribution* of portfolio losses, which reflects how the portfolio could behave given the historical changes in the risk factors. The method assumes the distribution of $L_{t+1}(\boldsymbol{X}_{t+1})$ is discrete and takes on each of the values $\tilde{L}_s$ with equal probability $1/n$ ([7]). To estimate VaR at a given confidence level $\alpha$, we calculate the *empirical quantile* of the simulated loss distribution. Suppose the values of the data are ordered by

$$\tilde{L}_{n,n} \leq \dots \leq \tilde{L}_{1,n}.$$

An estimator for VaR at confidence level $\alpha$ then can be the $(1-\alpha)$-quantile of the ordered losses. Specifically, $\mathrm{VaR}_\alpha$ is given by

$$\mathrm{VaR}_\alpha(L) = \tilde{L}_{[n(1-\alpha)],n},$$

where $[n(1-\alpha)]$ represents the largest integer not exceeding $n(1-\alpha)$. For example, if there are 1000 simulated losses (i.e.,$n = 1000$) and $\alpha = 0.99$, we would select the 10th largest loss (since $1000 \times (1-0.99) = 10$).

#### 2.2.4 Monte Carlo Simulation

The *Monte Carlo method* is a widely used approach in risk management that involves simulating an explicit parametric model for risk-factor changes. The first step of the method is to choose an appropriate model and calibrate it using historical risk-factor change data $\boldsymbol{X}_{t-n+1}, \dots, \boldsymbol{X}_t$. Once the model is calibrated, it is used to generate multiple independent realizations of future risk-factor changes, denoted as $\tilde{\boldsymbol{X}}_{t+1}^{(1)}, \dots, \tilde{\boldsymbol{X}}_{t+1}^{(m)}$, where $m$ represents the number of independent realizations simulated.
In a similar fashion to the historical simulation method, the loss operator is applied to these simulated risk-factor vectors to calculate simulated portfolio losses. Specifically, for each simulation, the portfolio loss is calculated using the loss operator applied to each simulated risk-factor change. These simulated losses are then used to estimate VaR.
The number of replications, $m$, can be chosen freely, allowing for more accurate estimates of risk measures compared to historical simulation, especially when a large number of simulations is used.

#### 2.2.5 Comparison of VaR Methods

The three main methods for calculating Value-at-Risk, variance-covariance, historical simulation, and Monte Carlo simulation each have their strengths and weaknesses, as explained by McNeil et al in ([1], Chapter.2.3) and Jorion in ([8], Chapter 10).

The variance-covariance method is easy to use and quick to compute. It provides a simple formula for calculating risk using a covariance matrix, which makes it popular for straightforward risk management. However, it relies on two big assumptions that can make it less accurate. First, it assumes the relationship between risk factors and portfolio returns is linear, which might not always be true. Second, it assumes that risk factors follow a normal distribution, which tends to underestimate the chance of extreme losses. This method also struggles to handle portfolios with nonlinear instruments like options or mortgages.

The historical simulation method is simple and intuitive because it uses actual historical data rather than making assumptions about how risk factors behave. This makes it effective at capturing extreme events in the data and useful for explaining past risks. However, its accuracy depends on having good quality historical data. If the data window is too short, it might miss important events, while a long window may include outdated scenarios. It also assumes that past patterns will repeat in the future, which may not always happen. Additionally, the method can be less precise for high confidence levels and small sample size.

The Monte Carlo simulation method is the most powerful and flexible approach. It is excellent for complex portfolios because it can handle issues like changing volatility, extreme scenarios, and nonlinear instruments. However, this method requires a lot of computational power and expertise to implement. It can also be slow and expensive because it involves running many simulations. Additionally, the results depend on the assumptions made about how risk factors behave, which introduces the risk of errors if the model is wrong. The accuracy of the method also depends on running a large number of simulations, which can increase variability in the results if not done properly.

## 2.3 VaR with Time-dependent Volatility

In Section 2.2, we introduced methods for calculating Value-at-Risk that rely on analyzing the unconditional loss distribution. This approach assumes that the risk factors are drawn from a stationary process, meaning that past data is used as a reliable indicator of future risk ([1], p.49).
However, in real-life financial markets, volatility is not consistent over time. Volatility tends to vary based on market conditions: it may be low during stable periods but increase significantly during market shocks or crises ([3], p.201). Therefore, relying on an unconditional distribution that assumes constant volatility may not accurately reflect the changing nature of financial markets.
In this section, we will introduce models that take into account the changes in volatility over time. In our work, we will focus on two such models: Generalized Autoregressive Conditional Heteroskedasticity (GARCH) and a special case, Exponentially Weighted Moving Average (EWMA). The same approach for describing these two methods, as outlined in Chapter 14.3 of [2], is considered with some explanation from Chapter 10.6 and Chapter 10.7 of [3].

### 2.3.1 Generalized Autoregressive Conditional Heteroskedasticity (GARCH)

The *generalized autoregressive conditional heteroskedasticity* (GARCH) process is an econometric model developed by Engle (1982) and Bollerslev (1986), which describes an approach to estimate volatility in financial markets ([2], p. 340).
Consider a dataset of daily return values, $X_1, \ldots, X_n$, derived from the logarithmic differences of a price, index, or exchange rate series $(S_t)_{t=0,1,\ldots,n}$. These returns are calculated as

$$X_t = \ln(S_t/S_{t-1}). \tag{7}$$

The GARCH model assumes that log-return $(X_t)$ at time $t$, has a normal distribution conditional on the mean $\mu_t$ and the standard deviation $\sigma_t$, that is, $X_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$.
The model focuses on the conditional variance, $h_t = \sigma_t^2$, which depends on current and past information. Unlike simple models, here the standard deviation $\sigma_t$ changes over time (time-varying volatility). This contrasts with the unconditional variance, which remains constant over time. The GARCH(1,1) model describes the conditional variance as

$$h_t = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta h_{t-1},$$

where $\alpha_0 > 0$ is the constant term, $\alpha_1 > 0$ represents the influence of the most recent squared return $X_{t-1}^2$, and $\beta > 0$ accounts for the persistence of past conditional variance $h_{t-1}$ .
The stationarity of the model requires $\alpha_1 + \beta < 1$, ensuring the variance process stabilizes at a long-term average. The unconditional variance $h$ is given by

$$h = \frac{\alpha_0}{1 - \alpha_1 - \beta}.$$

The GARCH model can also be used to forecast future variances. For the next-day forecast, the variance forecast depends on today's variance and the parameters of the model

$$E_{t-1}(X_{t+1}^2) = \alpha_0 + \alpha_1 E_{t-1}(X_t^2) + \beta h_t = \alpha_0 + \gamma h_t,$$

where $\gamma = \alpha_1 + \beta$.
For multi-step forecasting, this recursive structure propagates, allowing for variance forecasts to stabilize toward the long-run variance as time horizons extend.

### 2.3.2 Exponentially Weighted Moving Average (EWMA)

The *Exponentially Weighted Moving Average* (EWMA) model is a simplified version of the GARCH model, with a particular structure that emphasizes recent observations while exponentially reducing the weight of older data ([2], p.342). In the EWMA framework, the variance forecast is based on a weighted average of the previous forecast and the latest squared return. Mathematically, it is expressed as

$$h_t = \lambda h_{t-1} + (1 - \lambda) X_{t-1}^2, \tag{8}$$

where $h_t$ is the conditional variance at time $t$, $\lambda$ is the decay factor, which determines the relative weight of past observations and $X_{t-1}^2$ is the squared return at time $t - 1$.
The EWMA model applies exponentially decreasing weights to past squared returns. This means that recent returns have the most influence on the forecast, while older returns gradually contribute less.
To gain a clearer understanding of why equation (8) represents exponentially decreasing weights, we begin by substituting $h_{t-1}$ into the equation. This leads to the following expression

$$h_t = \lambda[\lambda h_{t-2} + (1 - \lambda) X_{t-2}^2] + (1 - \lambda) X_{t-1}^2.$$

This can also be rewritten as

$$h_t = (1 - \lambda)(X_{t-1}^2 + \lambda X_{t-2}^2) + \lambda^2 h_{t-2}.$$

By substituting $h_{t-2}$ in a similar manner, we obtain

$$h_t = (1 - \lambda)(X_{t-1}^2 + \lambda X_{t-2}^2 + \lambda^2 X_{t-3}^2) + \lambda^3 h_{t-3}.$$

And repeating this process shows that

$$h_t = (1 - \lambda) \sum_{i=1}^{m} \lambda^{i-1} X_{t-i}^2 + \lambda^m h_{t-m}, \quad m \in \mathbb{N}.$$

For a larger dataset, the term $\lambda^m h_{t-m}$ is sufficiently small to be ignored from the equation. The total variance forecast can be rewritten as

$$h_t = (1 - \lambda) \left[ X_{t-1}^2 + \lambda X_{t-2}^2 + \lambda^2 X_{t-3}^2 + \dots \right].$$

Here, the weight on the return $X_{t-k}^2$ is $(1 - \lambda)\lambda^{k-1}$. This shows that as $k$ increases, the weight diminishes exponentially, meaning older observations contribute less and less to the variance forecast.

The EWMA approach is efficient as it requires minimal data storage, needing only the prior variance estimate and the latest observation. The decay factor $\lambda$ serves as the smoothing parameter. A higher $\lambda$ value (close to one) implies slower decay in the series, retaining more data points that "fall off" gradually. Conversely, a lower $\lambda$ indicates faster decay, where the weights decrease more rapidly, resulting in fewer data points being used due to the rapid drop-off. The RiskMetrics database applies EWMA with $\lambda = 0.94$ for daily volatility updates.

The EWMA model is a special case of GARCH(1,1) where $\alpha_0 = 0$ and $\alpha_1 = 1 - \lambda$ and $\beta = \lambda$, meaning the EWMA model has permanent persistence ($\alpha_1 + \beta = 1$) and ignores any impact on long run variance. This makes the model particularly useful for short-term forecasting, where capturing the immediate impact of recent observations is more important.

**Remark 2.2.** Both EWMA and GARCH provide the conditional volatility $\sigma_t$ that can be used in parametric VaR calculations. After forecasting volatility using these models, we can use this volatility estimate to calculate VaR. In the variance covariance method we can use the formula from Lemma 2.1 when assuming a normal distribution. As for the student's t-distribution, we can use the formula from Lemma 2.2.

**Remark 2.3.** There are several approaches to improve historical simulation (HS) described in section 2.2.3. One way is by using the EWMA model, as suggested by Boudoukh et al.[9]. Instead of giving equal weight to all past data, we apply decreasing weights to older returns. Then, we select the desired percentile from this weighted distribution. This approach is called hybrid historical simulation or age-weighted historical simulation. Other methods, such as volatility-weighted historical simulation and filtered historical simulation, are discussed by K. Dowd in Chapter 4.4 of [10].

### 2.4 Backtesting Value-at-Risk

In sections 2.2 and 2.3 different methods for calculating VaR were explained. However, these methods and VaR itself have several limitations that make the accuracy of risk estimates questionable. VaR models are only valuable if they can accurately predict future risks. To ensure the estimates are reliable, it is important to backtest the models using proper methods ([11], p.16).

Backtesting is a statistical procedure used to verify the accuracy and reliability of Value-at-Risk models by comparing actual financial outcomes with the model's projections. It acts as a "reality check" to determine whether the VaR estimates accurately capture market risk. If the results indicate discrepancies, this may signal issues such as incorrect assumptions, faulty modeling, or improper parameter settings ([8], p.153).

Exceptions (or exceedances) occur when observed losses exceed the model's predicted VaR. For a well-calibrated model, the number of exceptions should be in line with the confidence level, for example, around 1% for a 99% confidence level. Too many exceptions mean the model underestimates risk, while too few exceptions indicate the model overestimates risk([8], 154).

Several methods have been developed to backtest VaR. A thorough review of these backtesting techniques is provided in [12], offering a detailed overview of the available approaches. In this section, we will narrow our focus to two key tests: Kupiec's proportion of failures (POF) test, and the Christoffersen test.

### 2.4.1 Kupiec Test

The process of backtesting VaR, involves assessing the accuracy of a VaR model by comparing predicted losses to actual losses. The most common approach is to record *the failure rate* $x/n$, where $x$ is the number of exceptions and $n$ is the total number of observations. This proportion should ideally converge to $p := 1 - \alpha$, the probability level corresponding to the confidence level as the sample size increases. For instance, a 99% confidence level implies that violations should occur approximately 1% of the time. It is uncommon to observe the exact number of exceptions implied by the confidence level. Therefore, statistical analysis is required to determine whether the observed number of exceptions is within a reasonable range, ultimately deciding whether the model should be accepted or rejected ([11], p.17).

Under the *null hypothesis* that the model is correct, the occurrence of exceptions follows a sequence of independent *Bernoulli trials*, where each trial has two possible outcomes: either a VaR violation (success) or no violation (failure). Consequently, the number of exceptions $x$ follows a *binomial distribution*

$$P(x = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, 3, ...., n$, where $x$ has the expected value $E(x) = pn$ and variance $V(x) = p(1-p)n$. For large sample sizes $n$, the binomial distribution can be approximated by a normal distribution. The standardized form of the test statistic becomes

$$z = \frac{x - pn}{\sqrt{p(1-p)n}},$$

which is approximately $\sim N(0, 1)$.
To statistically test whether the observed failure rate $\hat{p} = x/n$ significantly deviates from the expected failure rate $p$, Kupiec (1995) proposed the proportion of failures (POF) test. This test evaluates the null hypothesis $H_0 : p = \hat{p}$ using a likelihood-ratio (LR) test statistic defined as

$$LR_{\text{POF}} = -2 \ln \left( \frac{(1-p)^{n-x} p^x}{\left(1 - \left(\frac{x}{n}\right)\right)^{n-x} \left(\frac{x}{n}\right)^x} \right). \tag{9}$$

Under the null hypothesis, $LR_{\text{POF}}$ follows a chi-squared distribution with one degree of freedom. If $LR_{\text{POF}}$ exceeds the critical value of the chi-squared distribution, the null hypothesis is rejected, and the model is considered inaccurate.

A critical consideration in backtesting is the tradeoff between *Type 1* and *Type 2 errors*. A type 1 error occurs when a correct model is rejected, whereas a type 2 error happens when an incorrect model is not rejected. A powerful test minimizes both errors, i.e., avoiding the rejection of good models (type 1) while ensuring that bad models are not accepted (type 2). Jorion illustrates the concepts in a straightforward way in Chapter 6.2 in [8].

Kupiec introduces approximate 95% confidence regions for such test, based on the tail values of the *log-likelihood ratio*, which follows a chi-squared distribution with one degree of freedom as the sample size $n$ increases. The null hypothesis is rejected if the test statistic $LR_{\text{POF}} > 3.841$, (the critical value for the chi-squared distribution at a 95% confidence level)([8], p.160).

Table 1 summarizes these regions which define the range of acceptable exceptions for different confidence levels and sample sizes, ensuring the model's performance is evaluated against statistically significant thresholds.

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

Table 1: Nonrejection regions for number of failures $N$ for POF-test with different sample sizes and confidence levels as presented by Kupiec in [13].

Assuming $\alpha = 0.99$, with one year of data ($n = 252$ trading days), we would expect to observe approximately $N = pn = 1\% \times 252$ exceptions. However, as long as the number of exceptions $N$ remains below 7 ($N < 7$), the null hypothesis cannot be rejected. If $N$ reaches 7 or higher ($N \geq 7$), it suggests that the Value-at-Risk is too low or that the model underestimates the likelihood of large losses.
As the sample size increases, the confidence intervals for the POF-test become narrower, making it easier to reject an incorrect model. For example, at the 95% confidence level with 252 observations, the interval for $\frac{x}{n}$ is $[6/252 = 0.024,\ 20/252 = 0.079]$. With a larger sample size of 1000 observations, the interval becomes much smaller given the values $[37/1000 = 0.037,\ 65/1000 = 0.065]$. This shows that increasing the sample size reduces the range of acceptable values ([8], p.161).

**Remark 2.4.** Note that N represents the number of failures that can occur in a sample of size $n$ without rejecting the null hypothesis that $p$ is the correct probability, based on a 95% confidence level for the backtest. It is also important to highlight that the confidence level used for backtesting is not the same as the confidence level used when calculating the actual VaR.

### 2.4.2 Christoffersen Test

Christoffersen's (1998) *interval forecast test* is an extension of Kupiec's POF-test, designed to assess both the correct coverage and the independence of exceptions in VaR models. While Kupiec's test focuses solely on whether the frequency of violations is in line with the expected failure rate (unconditional coverage), Christoffersen introduces a framework to also test if exceptions occur *independently* over time, i.e., the test checks if a VaR violation on one day is affected by what happened the day before. ([11], p.27).

The approach to this test is based on the methodology presented by Olli Nieppola in ([11], p.27). Following this framework, consider portfolio return data for $n$ days. For each day, we define an indicator variable $I_t$, which equals 1 if a VaR violation occurs (i.e., the portfolio loss exceeds the VaR) and 0 otherwise

$$I_t = \begin{cases} 1 & \text{if a VaR violation occurs} \\ 0 & \text{if no VaR violation occurs} \end{cases}$$

For $i, j \in \{0, 1\}$, let $n_{i,j}$ represent the number of days where state $j$ occurs, given that state $i$ occurred on the previous day. The independence test checks whether the probability of observing an exception today depends on whether an exception occurred on the previous day. This is done by defining a $2 \times 2$ contingency table that counts the transitions between states $I_{t-1}$ (previous day) and $I_t$ (current day). The table is structured as follows

| | $I_{t-1}=0$ | $I_{t-1}=1$ | Total |
|---|---|---|---|
| $I_t=0$ | $n_{00}$ | $n_{10}$ | $n_{00}+n_{10}$ |
| $I_t=1$ | $n_{01}$ | $n_{11}$ | $n_{01}+n_{11}$ |
| Total | $n_{00}+n_{01}$ | $n_{10}+n_{11}$ | $n$ |

From this table, we calculate the conditional probabilities

$$\pi_0 = \frac{n_{01}}{n_{00}+n_{01}} \quad \text{and} \quad \pi_1 = \frac{n_{11}}{n_{10}+n_{11}},$$

where $\pi_0$ is the probability of observing a violation when the previous day had no violation ($I_{t-1}=0$), and $\pi_1$ is the probability of a violation following a violation on the previous day ($I_{t-1}=1$). The overall probability of observing a violation is

$$\pi = \frac{n_{01}+n_{11}}{n},$$

where $n$ is the total number of observations.
Under the null hypothesis of *independence* $H_0 : \pi_0 = \pi_1$, the likelihood ratio statistic is given by

$$LR_{\text{ind}} = -2\ln\left(\frac{(1-\pi)^{n_{00}+n_{10}}\pi^{n_{01}+n_{11}}}{(1-\pi_0)^{n_{00}}\pi_0^{n_{01}}(1-\pi_1)^{n_{10}}\pi_1^{n_{11}}}\right).$$

This statistic follows a $\chi^2$-distribution with one degree of freedom. If $LR_{\text{ind}}$ exceeds the critical value, we reject the null hypothesis, indicating that exceptions are not independent and may exhibit clustering over time.
To test *conditional coverage* (both correct coverage and independence), Christoffersen combines the independence test with Kupiec's POF-test. The joint likelihood ratio statistic is

$$LR_{\text{cc}} = LR_{\text{POF}} + LR_{\text{ind}},$$

where $LR_{\text{POF}}$ tests unconditional coverage, and $LR_{\text{ind}}$ tests independence of exceptions.
The combined test statistic $LR_{\text{cc}}$ follows a $\chi^2$-distribution with two degrees of freedom. If $LR_{\text{cc}}$ exceeds the critical value, the VaR model is rejected for failing either correct coverage, independence, or both.
A joint test may sometimes pass, while the individual components coverage or independence fail. Therefore, it is advisable to separately evaluate $LR_{\text{POF}}$ and $LR_{\text{ind}}$ to identify the source of the problem. In this case each statistic, $LR_{\text{POF}}$ and $LR_{\text{ind}}$, is calculated separately, with the chi-squared distribution with one degree of freedom serving as the critical value for both statistics ([11], p.28). Christoffersen's framework provides a robust method for assessing both the frequency and independence of VaR exceptions.

# 3 Methodology

## 3.1 Data

The dataset used in this thesis consists of the stocks that make up the Standard & Poor's 500 (S&P 500) index. The study spans ten years, starting on December 1, 2014, and ending on December 1, 2024 with 2516 observations.
The daily returns are calculated using adjusted closing prices, which take into account stock splits, dividends, and other changes. This method ensures that the data reflects the true changes in stock value. Continuous returns are used using the formula 7 in section 2.3.1. The data was obtained from Yahoo Finance[1] using the `tidyquant` package in R.

Some of the descriptive statistics for the daily log returns of the S&P 500 index are shown in Table 2. The dataset contains 2,516 observations with no missing values. The returns range from a minimum of -0.1277 to a maximum of 0.0897. The central measures indicate a mean daily return of 0.0004, with the first quartile (Q1) at -0.0038, the median at 0.0007, and the third quartile (Q3) at 0.0057. The standard deviation, representing volatility, is 0.0113. The distribution exhibits negative skewness (-0.8019), indicating more frequent small positive returns and fewer extreme negative returns. Additionally, the kurtosis is high (15.6966), reflecting the presence of fat tails and a greater likelihood of extreme outcomes, common in financial return data.

| **Statistic** | **Value** |
|---|---|
| Number of Observations | 2,516 |
| Missing Values (NAs) | 0.0000 |
| Minimum | -0.1277 |
| Quartile 1 (Q1) | -0.0038 |
| Mean | 0.0004 |
| Median | 0.0007 |
| Quartile 3 (Q3) | 0.0057 |
| Maximum | 0.0897 |
| Standard Deviation | 0.0113 |
| Skewness | -0.8019 |
| Kurtosis | 15.6966 |

Table 2: Descriptive statistics for S&P 500 daily log returns (December 2014- December 2024).

Figures 3 and 4 show the adjusted prices and the corresponding daily returns derived from these prices, respectively. Figure 3 shows that the S&P 500 adjusted closing prices rose steadily over the ten years, with a sharp drop in early 2020 during the COVID-19 pandemic. Despite this, the index quickly recovered and reached new highs by 2024. While Figure 4 shows the daily returns moving around zero, with clear spikes in early 2020 due to market instability. After 2020, the returns became more stable, though some fluctuations still occurred, showing the market's changing nature.

[1] The data is available on http://finance.yahoo.com/.


Figure 3: S&P500 adjusted closing prices (December 2014- December 2024).

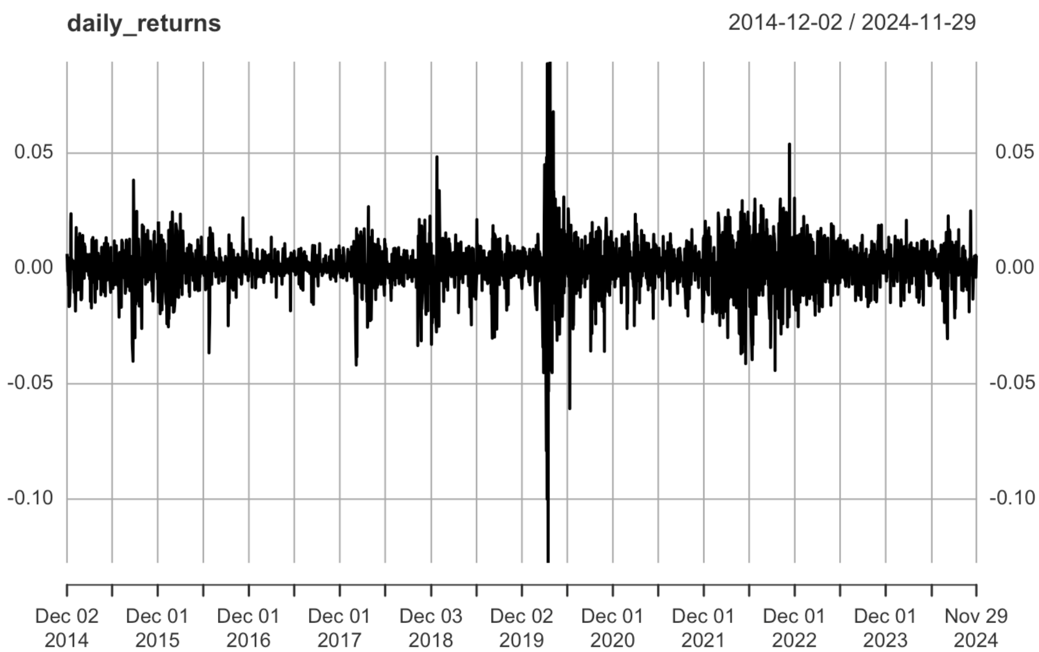



Figure 4: S&P500 daily returns (December 2014- December 2024).

**Normality Check**

Before implementing the VaR forecasting methods, the dataset was checked for normality. This step was crucial as many VaR estimation methods assume a normal distribution of returns. Deviations from normality could affect the accuracy and reliability of the methods.

To evaluate the normality of daily returns for the S&P 500 over the last 10 years, we conducted a Shapiro-Wilk test, a widely used statistical test for assessing the normality of a dataset [14]. The test yielded a test statistic of $W = 0.87695$ and a p-value $< 2.2 \times 10^{-16}$. This result strongly suggests the rejection of the null hypothesis that the returns are normally distributed.

The Q-Q plot shown in Figure 5 further supports this conclusion, as the sample quantiles (black points) deviate significantly from the reference line (red line), particularly in the tails where the points curve away from the line. This deviation indicates the presence of outliers in the tails, suggesting extreme values that may compromise the reliability of the variance-covariance method and GARCH models with normal innovations. Consequently, results based on this method should be interpreted with caution.
Additionally, the skewness of $-0.8019$ highlights a pronounced left skew in the data, while the kurtosis of 15.6966 underscores the heavy-tailed nature of the returns, as shown in Table 2. Despite these deviations from normality, the normal distribution remains a useful approximation in certain situations. As highlighted in [15] by Costa et al., while financial returns often deviate from normality, the normal distribution can still provide practical insights for tools like Value at Risk (VaR). However, its limitations, particularly in underestimating tail risks, must be carefully considered in risk-sensitive scenarios.

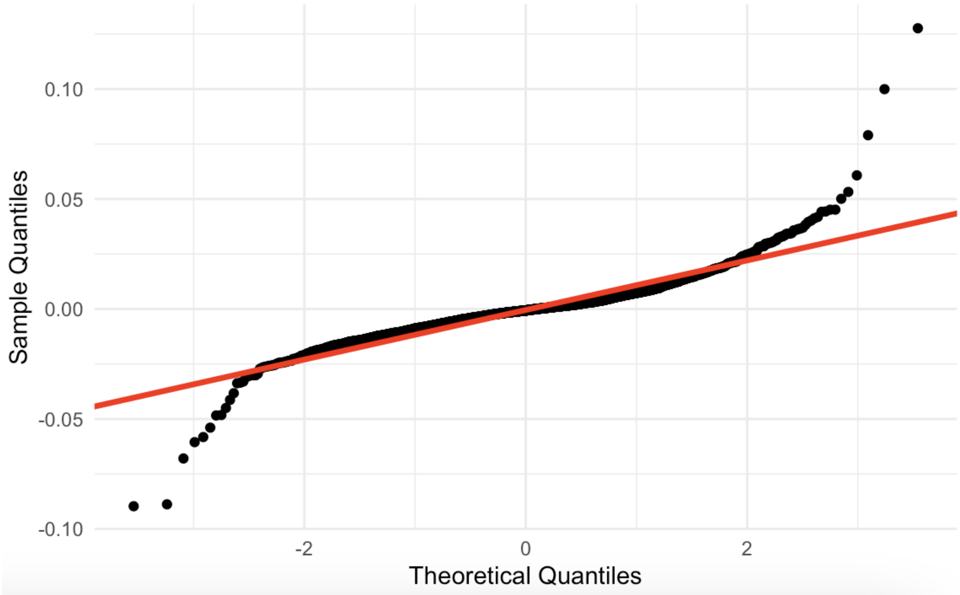



Figure 5: Q-Q Plot of Daily Returns for S&P 500.

## 3.2 Forecasting VaR Methods

The empirical part of this thesis will focus on applying the theoretical methods described in sections 2.2 and 2.3 to actual financial data, demonstrating the calculation, improvement, and backtesting of Value-at-Risk. Specifically, we will estimate VaR using historical simulation and variance-covariance methods. We will enhance the accuracy of these methods by incorporating conditional volatility models such as *exponentially weighted moving average* (EWMA) with decay factor $\lambda = 0.94$[2] and *generalized autoregressive conditional heteroskedasticity* (GARCH). To assess the validity of these VaR estimates, backtesting will be conducted using the Kupiec test and the Christoffersen test, ensuring that the results meet statistical and practical accuracy standards.

Our analysis is inspired by the example provided in the book Quantitative Risk Management by McNeil et al ([1], p.55). In their study, they evaluate 95% and 99% VaR estimates for a portfolio exposed to five risk factors: three international equity indices and two exchange rates. The methods employed include unconditional and conditional approaches, as well as backtesting

[2]The decay factor $\lambda = 0.94$ is commonly used in financial models (e.g., RiskMetrics) for daily returns as it effectively balances responsiveness to recent changes and smoothness in volatility estimation.

to verify their effectiveness. However, our analysis differs from McNeil's example in the following ways: McNeil' s example considers a portfolio influenced by multiple risk factors, including indices and exchange rates, while our study focuses solely on the SP 500 index, simplifying the analysis to a single risk factor. Additionally, McNeil's study spans the years 1996-2003, capturing events like the Dot-com bubble, whereas we will analyze data from December 2014 to December 2024, covering recent market dynamics such as the COVID-19 pandemic and its aftermath. Furthermore, McNeil includes an advanced method, HS-CONDEVT, which incorporates extreme value theory, but we exclude this method as it falls outside the scope of our study. Finally, while McNeil primarily evaluates violations, we will explicitly use Kupiec and Christoffersen tests to assess the accuracy of VaR estimates.

The analysis focuses on the S&P 500 index over a ten-year period. The portfolio consists of the S&P 500, making the portfolio loss directly proportional to the index's log-returns. The loss operator for this case is defined as $l^{\Delta}_{[t]}(x) = -x$, where $x$ represents the log-return of the S&P 500.
The goal is to estimate VaR at the 95% and 99% confidence levels for all trading days during 2019-2024. The estimation uses the last 1,000 days of historical data for risk-factor returns. To estimate the Value-at-Risk, we use the most recent 1,000 daily log-returns, following a rolling-window approach. Specifically, for each trading day $t+1$, VaR is estimated based on the historical data from days $t-999$ to $t$. This dynamic approach ensures that VaR estimates reflect the evolving market conditions. The analysis spans the period from December 1, 2018, to December 1, 2024, allowing for a sufficient warm-up period during the initial years (December 1, 2014, to November 30, 2018) to build a robust dataset for rolling-window calculations.

By employing this methodology, we aim to assess the effectiveness of various VaR estimation techniques, including historical simulation, variance-covariance methods, and conditional models like EWMA and GARCH, applied to the S&P 500 index.

The methods in Table 3 will be employed to calculate and analyze VaR.

| **Symbol** | **Method** |
|---|---|
| HS | An unconditional method that uses historical data to estimate the quantile of the loss distribution. |
| HS-EWMA | A conditional version of historical simulation that uses the EWMA method to estimate changing volatility. |
| HS-GARCH | A conditional historical simulation method where GARCH(1,1) models with Gaussian innovations estimate volatility. |
| HS-GARCH-$t$ | A conditional historical simulation method where GARCH(1,1) models assume $t$-distributed innovations. |
| VC | An unconditional variance-covariance method assuming univariate Gaussian risk-factor changes. |
| VC-EWMA | A conditional method similar to VC but with the EWMA approach to estimate time-varying variance. |
| VC-GARCH | A conditional variance-covariance method where GARCH(1,1) models estimate the conditional variance. |
| VC-$t$ | An unconditional variance-covariance method assuming a univariate $t$-distribution to model heavier tails. |
| VC-GARCH-$t$ | A conditional variance-covariance method where GARCH(1,1) models use $t$-distributed innovations. |

Table 3: Methods employed for VaR calculation.

The study will evaluate the effectiveness of traditional and enhanced VaR estimation methods on S&P 500 data over the past decade. By employing advanced conditional models and backtesting procedures, we aim to provide robust insights into the reliability of these methods under real-world market conditions.

### 3.3 Backtesting Methods

The accuracy of the Value-at-Risk estimates was checked using backtesting. This process used the Kupiec test and the Christoffersen test to see if the observed violations, where actual losses exceeded the VaR, matched their expected frequency and if these violations happened randomly.

The data used for backtesting included daily returns of the S&P 500 index and the VaR estimates from the methods listed in Table 3. VaR estimates were generated at both 95% and 99% confidence levels. However, backtesting for both 95% and 99% VaR was conducted at a 95% confidence level. A violation was counted whenever the actual daily loss exceeded the predicted VaR value for the corresponding confidence level.

The backtesting process followed several steps. First, a binary sequence of violations was created for each VaR method by comparing actual daily returns with the VaR estimates at 95% or 99% confidence levels. The Kupiec test was used to check if the percentage of observed violations matched the expected rate (5% for 95% VaR and 1% for 99% VaR). Additionally, the Christoffersen test ensured that the violations occurred independently, rather than in clusters.

To account for market changes, a rolling-window approach was employed. The VaR for each day was based on the most recent 1,000 daily returns. Backtesting covered the entire dataset from 2019 to 2024, with additional yearly analyses to assess performance during different market conditions.

Backtesting was implemented in `R` using the `rugarch` and `dplyr` packages. Automated scripts calculated test statistics for each method to ensure consistent results. Outputs included the expected and observed number of violations, likelihood ratio statistics for both unconditional and conditional coverage, p-values, and decisions on the null hypotheses. Results were summarized in tables to compare all methods at both 95% and 99% confidence levels. The results for 2020 were also analyzed to highlight performance changes over time.

## 4 Results

This section presents the results of forecasting and backtesting Value-at-Risk (VaR) using the methods described in section 3. The analysis focuses on evaluating how well these methods estimate VaR at 95% and 99% confidence levels and how accurately they capture market risk over the study period. The results are divided into two parts: forecasting results, which compare predicted VaR estimates to actual market returns, and backtesting results, which assess the reliability of these methods using statistical tests.

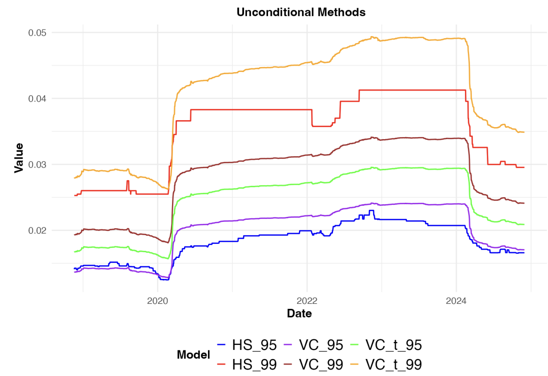
### 4.1 Forecasting Results

We will evaluate the performance of the methods for estimating Value-at-Risk at 95% and 99% confidence levels, focusing on their ability to capture market risk during the study period. Particular attention is given to periods of extreme volatility, such as the COVID-19 pandemic in early 2020.

### 4.1.1 Unconditional Methods

Our analysis begins with the unconditional methods: Historical Simulation (HS), Variance-Covariance (VC), and Variance-Covariance with $t$-distribution (VC-$t$). These methods rely entirely on historical data and assume constant market volatility, which limits their ability to adapt to sudden changes in risk. To assess their performance, we use two plots that compare predicted VaR values to daily returns and highlight the patterns of VaR estimates over time.


(a) Returns vs. VaR estimates for HS, VC, and VC-t at 95% and 99% confidence levels.



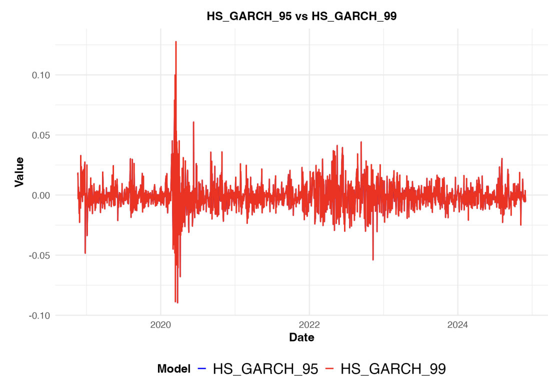

(b) VaR estimates for HS, VC, and VC-t over time at 95% and 99% confidence levels.

Figure 6: VaR forecasting results using unconditional methods: comparison of returns and predicted VaR estimates.

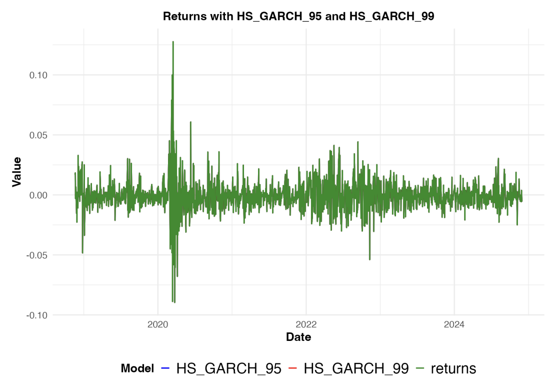
The first plot, shown in Figure 6(a) which overlays returns with the predicted VaR estimates, reveals that all three methods struggle during periods of extreme volatility. For example, during early 2020, returns frequently exceeded the predicted VaR valuess, particularly for the HS and VC methods, indicating that these approaches underestimated the level of market risk. The VC-$t$ method, on the other hand, consistently produced higher VaR estimates, which helped it better capture extreme losses. However, this came at the cost of overestimating risk during calmer periods, as its thresholds remained higher than necessary in less volatile conditions.
The second plot in Figure 6(b) provides further insights by showing how the methods' VaR estimates change over time. The HS method appears slow to react to shifts in market volatility, reflecting its reliance on historical returns without any adjustments for changing risk levels. The VC method, while providing smoother estimates, underestimates risk during periods of sharp volatility spikes, such as those in early 2020, as its assumption of a normal distribution fails to account for the fat tails present in financial return data. In contrast, the VC-$t$ method shows a larger increase in VaR estimates during volatile periods, reflecting its ability to account for extreme returns through its $t$-distribution assumption. However, this method's consistently higher estimates across the entire period suggest that it tends to be overly cautious even in stable markets.
Overall, the results highlight that while these methods share a reliance on historical data and constant volatility assumptions, their approaches lead to notable differences in performance. The HS method is slower to respond to changing conditions, the VC method underestimates risk during extreme events, and the VC-$t$ method is more responsive to tail risk but often overestimates risk in calmer markets. None of the three methods adapt dynamically to changes in volatility, which limits their ability to accurately estimate risk during rapidly changing market conditions. These findings underscore the need for more advanced methods, such as GARCH or EWMA, that can explicitly model time-varying volatility. Further validation through backtesting will be conducted to assess the accuracy of these methods in capturing real-world market risk.

### 4.1.2 HS-GARCH and HS-GARCH-t Methods

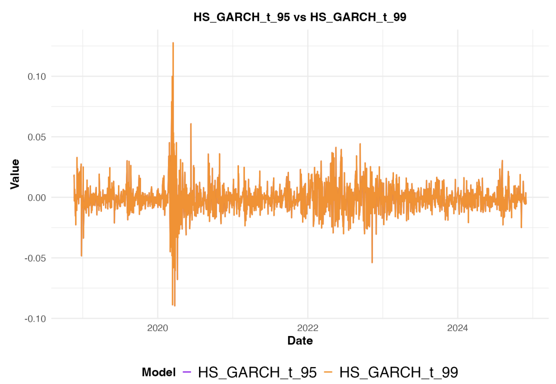
When examining the plots for the 95% VaR and 99% VaR generated using the HS-GARCH and HS-GARCH-t methods, some unusual patterns became apparent. For both methods, the 95% and 99% VaR estimates appear almost identical, which is unexpected as shown in Figure 7 (a) and (c). Typically, the 99% VaR estimates should lie significantly above the 95% VaR, as it represents a higher confidence level and, therefore, larger potential risk values.



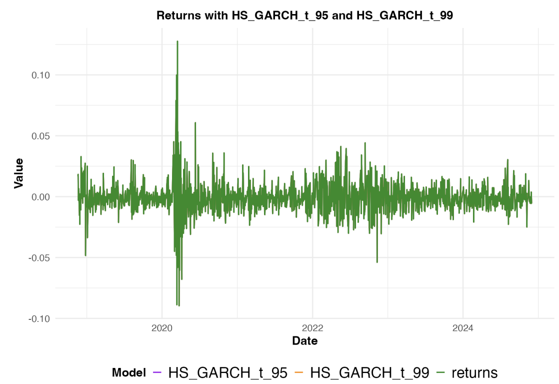

(a) VaR estimates for HS-GARCH at 95% and 99% confidence level.




(b) Returns vs. VaR estimates for HS-GARCH at 95% and 99% confidence level.




(c) VaR estimates for HS-GARCH-t at 95% and 99% confidence level.




(d) Returns vs. VaR estimates for HS-GARCH-t at 95% and 99% confidence level.

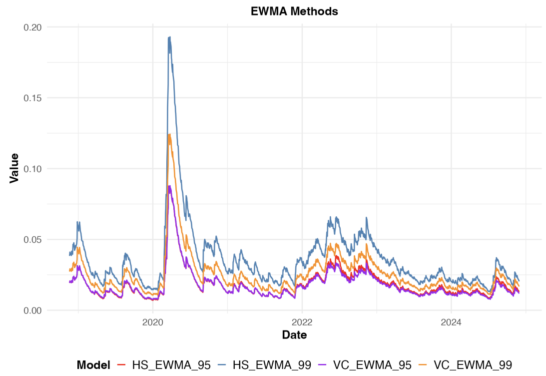
Figure 7: VaR forecasting results for HS-GARCH and HS-GARCH-t at 95% and 99% confidence level.

Additionally, as illustrated in Figure 7 (b) and (d), the VaR lines for these methods seem to closely mimic the daily returns, fluctuating in a way that resembles the return series rather than acting as stable thresholds. This behavior is unusual because VaR is supposed to represent a relatively steady boundary for extreme values, rather than reacting dynamically to short-term movements.

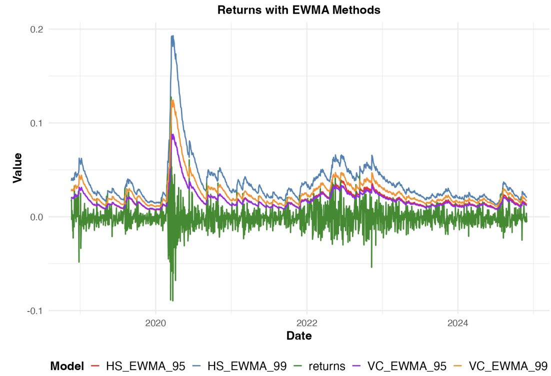
When comparing these four models-HS-GARCH 95%, HS-GARCH 99%, HS-GARCH-t 95%, and HS-GARCH-t 99%- to the other methods, only these stood out as being unusual. The plots for all other models showed expected behavior, with the 95% and 99% VaR lines clearly separated and providing smooth, logical estimates for risk. This strange behavior in the HS-GARCH and HS-GARCH-t methods will be explored further during the backtesting analysis later in the thesis, where their performance and validity will be assessed in greater detail.

### 4.1.3 EWMA Methods

The VaR estimates for both VC-EWMA and HS-EWMA methods are smooth and well-behaved. Unlike the previous methods (e.g., HS-GARCH), the plots clearly distinguish between the 95% and 99% VaR levels. Starting with Figure 8 (a), the VaR estimates for both methods behave as expected, with the 99% VaR consistently higher than the 95% VaR. This is logical since the 99% VaR accounts for a higher confidence level and reflects greater potential losses. Both methods show good responsiveness to market volatility, particularly during significant events like the 2020 COVID-19 crisis, where sharp spikes in VaR estimates are observed. These spikes gradually decline as market conditions stabilize.




(a) VaR estimates for EWMA models over time at 95% and 99% confidence levels.




(b) Returns vs. VaR estimates for EWMA models at 95% and 99% confidence levels.

Figure 8: VaR forecasting results using EWMA methods: comparison of returns and predicted VaR estimates.

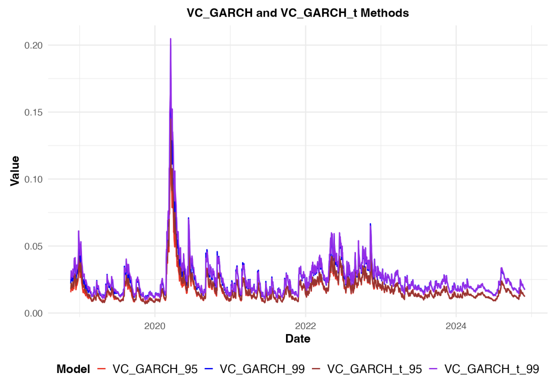
Focusing on the 95% VaR models specifically, the HS-EWMA and VC-EWMA methods behave almost the same way up to early 2022. Their lines closely overlap during this period, indicating they capture market dynamics similarly. However, after early 2022, the HS-EWMA line becomes more visible and diverges above the VC-EWMA line. This suggests that the HS-EWMA method begins estimating slightly higher risk compared to VC-EWMA. This difference occurs because HS-EWMA directly uses historical returns and reacts more strongly to past extreme losses, while VC-EWMA smooths these effects due to its reliance on normal distribution assumptions. As a result, HS-EWMA becomes more conservative in its risk estimates during this later period.

In Figure 8 (b), returns are plotted against the VaR estimates for both methods at 95% and 99% confidence levels. Both methods perform well, with the VaR estimates acting as effective boundaries for the returns. Breaches, where returns exceed the VaR values, are rare, which aligns with expectations for models at these confidence levels. The 99% VaR estimates are breached even less frequently than the 95% estimates, as expected. Comparing the two methods, HS-EWMA provides slightly higher estimates during volatile periods, making it more conservative in capturing tail risks. On the other hand, VC-EWMA provides smoother and slightly lower estimates, which may underestimate risks in extreme market conditions.

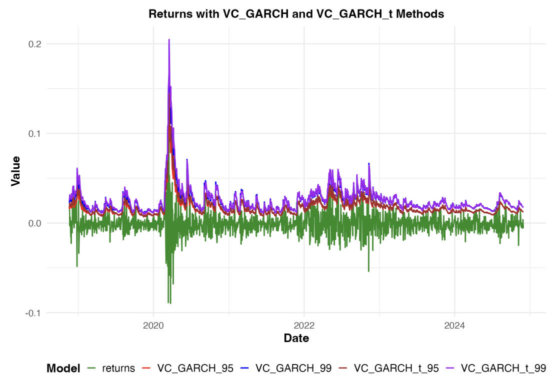
### 4.1.4 VC-GARCH and VC-GARCH-t Methods

The plots in Figure 9(a) show the performance of the VC-GARCH and VC-GARCH-t methods for estimating VaR at 95% and 99% confidence levels. Both methods produce smooth and well-behaved VaR estimates, similar to what was observed with the EWMA methods. The 99% VaR consistently lies above the 95% VaR, reflecting proper calibration for the two confidence levels. Both models respond effectively to market volatility, with significant spikes during the COVID-19 crisis in 2020 and a gradual stabilization afterward.

For the 95% and 99% VaR estimates, there are no significant differences between the two models. When plotted together, the lines for VC-GARCH and VC-GARCH-t at each confidence level are nearly identical. This similarity occurs despite VC-GARCH-t assuming a t-distribution for innovations, which is designed to better capture heavy tails. In this dataset, the impact of the heavier tails in the t-distribution appears negligible, resulting in almost identical VaR estimates. Because of this, separate plots for the two models were not included in the thesis to avoid redundancy and an overload of figures. However, the performance of these models and whether their similarity holds will be further investigated in the backtesting process.




(a) VaR estimates for VC-GARCH and VC-GARCH-t models at 95% and 99% confidence levels.




(b) Returns vs. VaR estimates for VC-GARCH and VC-GARCH-t models at 95% and 99% confidence levels.

Figure 9: VaR forecasting results using VC-GARCH and VC-GARCH-t methods: comparison of returns and predicted VaR estimates.

In Figure 9 (b), where returns are plotted alongside the VaR values, both models perform well. The VaR estimates effectively bound the returns, with only rare breaches, as expected for models calibrated at 95% and 99% confidence levels. The 99% VaR values are breached even less frequently than the 95% values, demonstrating the models' reliability. While VC-GARCH-t theoretically provides slightly more conservative risk estimates due to its t-distribution assumptions, this difference is not visible in the performance of the models with this dataset.

## 4.2 Backtesting Process

This section evaluates the performance of the VaR models using backtesting to assess their accuracy and reliability. The backtesting process is conducted for the entire dataset (2019-2024) and separately for the year 2020, a period of heightened market volatility. The analysis includes both the Kupiec test (Unconditional Coverage) and the Christoffersen test (Conditional Coverage). Models estimating VaR at both 95% and 99% confidence levels are tested at a 95% confidence level.

### 4.2.1 Backtesting for 95% VaR

The backtesting results presented here are for the entire period from 2019 to 2024. These results assess the accuracy of the VaR estimates for all methods using the Kupiec and Christoffersen tests. It is worth noting that the results for the HS_GARCH and HS_GARCH_t methods are not included in the tables. This will be explained later in the analysis.

Table 4 evaluates whether the actual number of violations, where losses exceed the VaR estimates, matches the expected number. For most methods, the Kupiec test fails to reject the null hypothesis ($H_0$), indicating that the observed violations are close to what is expected.

| Methods | Expected | Actual | UC LR Stat | UC Critical | UC P-Value | UC Decision |
|---|---|---|---|---|---|---|
| HS_95 | 75 | 82 | 0.51197 | 3.84146 | 0.47429 | Fail to Reject $H_0$ |
| HS_EWMA_95 | 75 | 82 | 0.51197 | 3.84146 | 0.47429 | Fail to Reject $H_0$ |
| VC_95 | 75 | 74 | 0.04787 | 3.84146 | 0.82682 | Fail to Reject $H_0$ |
| VC_EWMA_95 | 75 | 88 | 1.95265 | 3.84146 | 0.16230 | Fail to Reject $H_0$ |
| VC_GARCH_95 | 75 | 92 | 3.39934 | 3.84146 | 0.06522 | Fail to Reject $H_0$ |
| VC_t_95 | 75 | 53 | 8.06302 | 3.84146 | 0.00452 | Reject $H_0$ |
| VC_GARCH_t_95 | 75 | 86 | 1.37306 | 3.84146 | 0.24129 | Fail to Reject $H_0$ |

Table 4: Backtesting Results for 95% VaR - Unconditional Coverage, Kupiec test.

The HS_95 and HS_EWMA_95 methods both have 82 actual violations compared to the expected 75. This is slightly higher than expected, but the results are within acceptable limits, as the test fails to reject $H_0$. VC_95 performs particularly well, with 74 actual violations, very close to the expected 75. Its high p-value suggests strong alignment with the expected frequency. VC_EWMA_95 and VC_GARCH_95 show slightly higher actual violations (88 and 92, respectively), suggesting mild underestimation of risk. However, the test does not reject $H_0$, indicating these methods are still within acceptable limits. VC_t_95, on the other hand, significantly overestimates risk, with only 53 violations. The test strongly rejects $H_0$ ($p = 0.00452$), showing that this method fails to capture tail risk accurately. VC_GARCH_t_95 has 86 violations, slightly higher than expected, but still passes the test with $H_0$ not rejected.

Table 5 assesses not just the number of violations but whether they occur independently, rather than in clusters. This is a stricter test of the model's performance.

| Method | Expected | Actual | CC LR Stat | CC Critical | CC P-Value | CC Decision |
|---|---|---|---|---|---|---|
| HS_95 | 75 | 82 | 22.51324 | 5.99146 | 0.00001 | Reject $H_0$ |
| HS_EWMA_95 | 75 | 82 | 1.07386 | 5.99146 | 0.58454 | Fail to Reject $H_0$ |
| VC_95 | 75 | 74 | 17.55961 | 5.99146 | 0.00015 | Reject $H_0$ |
| VC_EWMA_95 | 75 | 88 | 1.95524 | 5.99146 | 0.37620 | Fail to Reject $H_0$ |
| VC_GARCH_95 | 75 | 92 | 5.52975 | 5.99146 | 0.06298 | Fail to Reject $H_0$ |
| VC_t_95 | 75 | 53 | 24.66605 | 5.99146 | 0.00000 | Reject $H_0$ |
| VC_GARCH_t_95 | 75 | 86 | 2.30023 | 5.99146 | 0.31660 | Fail to Reject $H_0$ |

Table 5: Backtesting Results for 95% VaR - Conditional Coverage, Christoffersen test.

HS_95 fails the Christoffersen test, as $H_0$ is rejected. This suggests that the violations for this method may occur in clusters. HS_EWMA_95, however, performs much better. The test fails to reject $H_0$, showing that violations are random and the method adapts well to market changes. VC_95, despite performing well in the Kupiec test, fails the Christoffersen test due to $H_0$ being rejected. This indicates clustering of violations and points to a limitation in handling dynamic market conditions. VC_EWMA_95 shows strong performance, with $H_0$ not rejected, meaning violations are well-distributed and occur randomly. This aligns with its adaptability seen in the earlier VaR plots. VC_GARCH_95 is close to the borderline, with $H_0$ not rejected ($p =$

0.06298), suggesting acceptable but slightly clustered violations. VC_t_95 continues to perform poorly, with $H_0$ rejected again. This confirms its inability to capture violations accurately, as noted in the Kupiec test. VC_GARCH_t_95 performs well, with $H_0$ not rejected, indicating good distribution and independence of violations.

From the results, HS_EWMA_95, VC_EWMA_95 and VC_GARCH_t_95 stand out as the best-performing methods, passing both tests and showing reliable estimates with random and well-distributed violations. VC_95, while accurate in terms of the number of violations, struggles with clustering, as shown in the Christoffersen test. The HS_EWMA_95 method performs well, especially compared to HS_95, which fails to handle violations independently. On the other hand, VC_t_95 consistently underperforms, failing to capture tail risk and struggling with both tests.
These backtesting results largely align with the trends observed in the earlier VaR plots. For example, HS_EWMA_95, VC_EWMA_95 and VC_GARCH_t_95, which showed stable and well-adapted VaR estimates, perform reliably here as well.

### 4.2.2 Backtesting for 99% VaR

The backtesting results for the 99% confidence level use the Kupiec and Christoffersen tests to evaluate the accuracy of the VaR estimates. The Kupiec test results, shown in Table 6, reveal mixed performance among the methods.

| Method | Expected | Actual | UC LR Stat | UC Critical | UC P-Value | UC Decision |
|---|---|---|---|---|---|---|
| HS_99 | 15 | 20 | 1.41205 | 3.84146 | 0.23472 | Fail to Reject $H_0$ |
| HS_EWMA_99 | 15 | 10 | 2.02308 | 3.84146 | 0.15492 | Fail to Reject $H_0$ |
| VC_99 | 15 | 37 | 22.63712 | 3.84146 | 0.00000 | Reject $H_0$ |
| VC_EWMA_99 | 15 | 37 | 22.63712 | 3.84146 | 0.00000 | Reject $H_0$ |
| VC_GARCH_99 | 15 | 35 | 19.12497 | 3.84146 | 0.00001 | Reject $H_0$ |
| VC_t_99 | 15 | 16 | 0.04506 | 3.84146 | 0.83189 | Fail to Reject $H_0$ |
| VC_GARCH_t_99 | 15 | 30 | 11.39958 | 3.84146 | 0.00073 | Reject $H_0$ |

Table 6: Backtesting Results for 99% VaR - Unconditional Coverage, Kupiec test.

The HS_99 method performs reasonably well, with 20 actual violations compared to the expected 15, and the test fails to reject the null hypothesis. Similarly, HS_EWMA_99 passes the test with only 10 violations, slightly overestimating risk but remaining statistically valid. In contrast, methods such as VC_99, VC_EWMA_99, and VC_GARCH_99 show significant deviations, with 35 to 37 actual violations compared to the expected 15, leading to a rejection of the null hypothesis and indicating underestimation of risk. The VC_t_99 method closely matches the expected number of violations (16 vs. 15), passing the test and showing good calibration. However, VC_GARCH_t_99, with 30 violations, fails the test, highlighting underestimation of risk.

The Christoffersen test results, in Table 7, indicate additional challenges for some methods. HS_99 struggles to pass this test, while HS_EWMA_99 performs strongly, highlighting its adaptability. Methods like VC_99, VC_EWMA_99, and VC_GARCH_99 continue to struggle, further confirming their difficulties in accurately estimating risk. VC_t_99, although passing the unconditional coverage test, shows challenges in the Christoffersen test, indicating that its violations may not occur independently. VC_GARCH_t_99 fails both tests, highlighting underestimation of risk and clustering of violations.
In summary, HS_EWMA_99 stands out as the most reliable method for 99% confidence level estimation, while other methods, particularly variance-covariance-based approaches, show significant limitations in their risk assessments. These findings align with trends seen earlier at

| Method | Expected Exceed | Actual Exceed | CC LR Stat | CC Critical | CC P-Value | CC Decision |
|---|---|---|---|---|---|---|
| HS_99 | 15 | 20 | 6.35867 | 5.99146 | 0.04161 | Reject $H_0$ |
| HS_EWMA_99 | 15 | 10 | 2.15589 | 5.99146 | 0.34029 | Fail to Reject $H_0$ |
| VC_99 | 15 | 37 | 28.90344 | 5.99146 | 0.00000 | Reject $H_0$ |
| VC_EWMA_99 | 15 | 37 | 22.64771 | 5.99146 | 0.00001 | Reject $H_0$ |
| VC_GARCH_99 | 15 | 35 | 20.77941 | 5.99146 | 0.00003 | Reject $H_0$ |
| VC_t_99 | 15 | 16 | 6.71321 | 5.99146 | 0.03485 | Reject $H_0$ |
| VC_GARCH_t_99 | 15 | 30 | 12.61097 | 5.99146 | 0.00183 | Reject $H_0$ |

Table 7: Backtesting Results for 99% VaR - Conditional Coverage, Christoffersen test.

the 95% level, with EWMA methods generally outperforming in adaptability.

### 4.2.3 Analysis of the Backtesting Results for 2020

To understand how the VaR models performed during a challenging year, we look at the backtesting results for 2020. This year was chosen because of its high market volatility and uncertainty. The analysis includes results from the Kupiec test (Unconditional Coverage) and the Christoffersen test (Conditional Coverage) at 95% confidence level. These results will also be compared to those from the full dataset (2019-2024) to see if there are any patterns or differences in how the models performed during this period.

For 95% VaR, methods like HS_95, VC_95, and VC_t_95 significantly underestimated the number of breaches, resulting in a rejection of the null hypothesis as shown in Table 8. These results indicate that these methods did not perform well in capturing risk during 2020. On the other hand, HS_EWMA_95, VC_EWMA_95, and VC_GARCH_t_95 performed much better, as the actual number of breaches aligned more closely with the expected values, leading to "Fail to Reject $H_0$" decisions.

| Method | Expected Exceed | Actual Exceed | UC LR Stat | UC Critical | UC P-Value | UC Decision |
|---|---|---|---|---|---|---|
| HS_95 | 12 | 32 | 22.29843 | 3.84146 | 0.00000 | Reject $H_0$ |
| HS_99 | 2 | 12 | 18.78315 | 3.84146 | 0.00001 | Reject $H_0$ |
| HS_EWMA_95 | 12 | 16 | 0.86472 | 3.84146 | 0.35242 | Fail to Reject $H_0$ |
| HS_EWMA_99 | 2 | 2 | 0.12083 | 3.84146 | 0.72813 | Fail to Reject $H_0$ |
| VC_95 | 12 | 28 | 14.79654 | 3.84146 | 0.00012 | Reject $H_0$ |
| VC_99 | 2 | 18 | 40.67328 | 3.84146 | 0.00000 | Reject $H_0$ |
| VC_EWMA_95 | 12 | 15 | 0.43484 | 3.84146 | 0.50962 | Fail to Reject $H_0$ |
| VC_EWMA_99 | 2 | 12 | 18.78315 | 3.84146 | 0.00001 | Reject $H_0$ |
| VC_GARCH_95 | 12 | 19 | 2.92697 | 3.84146 | 0.08711 | Fail to Reject $H_0$ |
| VC_GARCH_99 | 2 | 9 | 10.07068 | 3.84146 | 0.00151 | Reject $H_0$ |
| VC_t_95 | 12 | 23 | 7.25273 | 3.84146 | 0.00708 | Reject $H_0$ |
| VC_t_99 | 2 | 13 | 22.05887 | 3.84146 | 0.00000 | Reject $H_0$ |
| VC_GARCH_t_95 | 12 | 18 | 2.11770 | 3.84146 | 0.14560 | Fail to Reject $H_0$ |
| VC_GARCH_t_99 | 2 | 7 | 5.38792 | 3.84146 | 0.02028 | Reject $H_0$ |

Table 8: Backtesting Results for 2020 - Unconditional Coverage (UC), Kupiec test.

For 99% VaR, methods such as HS_99, VC_99, and VC_t_99 also struggled, showing a much higher number of breaches than expected. However, HS_EWMA_99 performed well, perfectly matching the expected breaches. VC_GARCH_t_99 showed slightly better results compared to other methods but still had issues, leading to a rejection of the null hypothesis.

The Analysis of the Christoffersen test results is presented in Table 9 indicates that for 95% VaR, methods such as HS_95, VC_95, and VC_t_95 showed evidence of clustering in the breaches, leading to a rejection of the null hypothesis. In contrast, methods like HS_EWMA_95, VC_EWMA_95, VC_GARCH_95 and VC_GARCH_t_95 showed no significant clustering, with decisions to "Fail to Reject $H_0$."

| Method | Expected Exceed | Actual Exceed | CC LR Stat | CC Critical | CC P-Value | CC Decision |
|---|---|---|---|---|---|---|
| HS_95 | 12 | 32 | 26.50943 | 5.99146 | 0.00000 | Reject $H_0$ |
| HS_99 | 2 | 12 | 21.31900 | 5.99146 | 0.00002 | Reject $H_0$ |
| HS_EWMA_95 | 12 | 16 | 3.96173 | 5.99146 | 0.13795 | Fail to Reject $H_0$ |
| HS_EWMA_99 | 2 | 2 | 0.15283 | 5.99146 | 0.92643 | Fail to Reject $H_0$ |
| VC_95 | 12 | 28 | 16.07545 | 5.99146 | 0.00032 | Reject $H_0$ |
| VC_99 | 2 | 18 | 42.70631 | 5.99146 | 0.00000 | Reject $H_0$ |
| VC_EWMA_95 | 12 | 15 | 1.63046 | 5.99146 | 0.44254 | Fail to Reject $H_0$ |
| VC_EWMA_99 | 2 | 12 | 19.07859 | 5.99146 | 0.00007 | Reject $H_0$ |
| VC_GARCH_95 | 12 | 19 | 4.52683 | 5.99146 | 0.10399 | Fail to Reject $H_0$ |
| VC_GARCH_99 | 2 | 9 | 10.73750 | 5.99146 | 0.00466 | Reject $H_0$ |
| VC_t_95 | 12 | 23 | 11.01545 | 5.99146 | 0.00406 | Reject $H_0$ |
| VC_t_99 | 2 | 13 | 24.07620 | 5.99146 | 0.00001 | Reject $H_0$ |
| VC_GARCH_t_95 | 12 | 18 | 4.15074 | 5.99146 | 0.12551 | Fail to Reject $H_0$ |
| VC_GARCH_t_99 | 2 | 7 | 5.78798 | 5.99146 | 0.05535 | Fail to Reject $H_0$ |

Table 9: Backtesting Results for 2020 - Conditional Coverage (CC), Christoffersen test.

For 99% VaR, methods like HS_99, VC_99, and VC_t_99 struggled again, with clear evidence of clustered breaches. HS_EWMA_99 stood out as the only method that performed well, while VC_GARCH_t_99 showed marginal results, with a p-value close to the threshold.

The results for 2020 highlight the strengths of HS_EWMA method, which consistently performed well across both tests and confidence levels. VC_EWMA also performed well for 95% VaR but struggled for 99% VaR for the Christoffersen test, where was rejected. In contrast, methods such as HS_95, VC_95, and VC_t_95 showed significant issues in capturing risk accurately, both in terms of the number of breaches and their independence. This suggests that EWMA-based methods were more effective in handling the volatility and dynamics of 2020. When compared to the results from the entire period (2019-2024), similar patterns are observed. Methods like HS_EWMA and VC_EWMA_95 consistently performed better in both datasets, while HS_95, VC_95, and VC_t_95 showed issues with overestimating breaches and clustering in both cases. However, the problems with methods such as VC_99 and VC_t_99 were more pronounced in 2020, likely due to the heightened volatility during that year. This indicates that while some methods are robust across different periods, others struggle more in years with extreme market conditions.

#### 4.2.4 Backtesting Issues with HS_GARCH Models

During the VaR estimation process for the HS_GARCH and HS_GARCH_t methods at both the 95% and 99% confidence levels, unusual behavior was observed in the VaR plots for these methods. The plots appeared strange and did not align with the expected patterns of other methods, as shown in Figure 7. To investigate this further, backtesting was conducted to determine whether this behavior would also be reflected in the statistical tests.
As we mentioned before in the previous section, backtesting was carried out using automated scripts in `R`. These scripts calculated the test statistics for the Kupiec and Christoffersen likelihood ratio (LR) tests, including the expected and observed number of violations, p-values and decisions on the null hypotheses. However, the results for the HS_GARCH and HS_GARCH_t methods were missing from the output tables, which required further investigation to understand the cause of this issue.
After doing some research [16], we found that the problem was likely due to the *number of observed breaches being zero* for these methods. Since these numbers were not calculated along with the other methods, we calculated the breaches separately to confirm this. The recalculation showed that the observed breaches for HS_GARCH and HS_GARCH_t were indeed zero. Table 10 summarizes the observed breaches for all methods:
The Kupiec and Christoffersen backtesting methods rely on likelihood ratio statistics, which

| **Model** | **Breaches** |
|---|---|
| HS_95 | 82 |
| HS_99 | 20 |
| HS_EWMA_95 | 82 |
| HS_EWMA_99 | 10 |
| HS_GARCH_95 | 0 |
| HS_GARCH_99 | 0 |
| HS_GARCH_t_95 | 0 |
| HS_GARCH_t_99 | 0 |
| VC_95 | 74 |
| VC_99 | 37 |
| VC_EWMA_95 | 88 |
| VC_EWMA_99 | 37 |
| VC_GARCH_95 | 92 |
| VC_GARCH_99 | 35 |
| VC_t_95 | 53 |
| VC_t_99 | 16 |
| VC_GARCH_t_95 | 86 |
| VC_GARCH_t_99 | 30 |

Table 10: Summary of observed breaches for all methods.

depend on the number of observed violations. Specifically, the *likelihood ratio (LR) statistic* for the Kupiec test is calculated as in equation (9).
When $x = 0$, the formula encounters issues because $\ln(0)$ and $0^0$ are undefined. This is the primary reason why backtesting does not work for methods with zero observed violations, as the calculation of $LR_{\text{POF}}$ becomes impossible.
The Christoffersen test, which builds on the Kupiec test by also checking the independence of violations over time, has a similar limitation. This test requires a binary sequence of violations to calculate whether they occur randomly or in clusters. However, when there are no observed violations, the binary sequence is empty, and the test cannot proceed.

The issue with the HS_GARCH and HS_GARCH_t methods seems to come from a few different factors. One possible reason is how the GARCH-estimated volatilities are combined with historical returns in the historical simulation framework. The GARCH and EWMA functions seem to work fine in other methods like VC_GARCH or HS_EWMA, so the issue is likely specific to how the historical simulation framework handles the scaling of returns with volatilities. The data itself could also play a role. Factors such as outliers, changes in volatility over time, or long periods of low volatility might exacerbate the problem, affecting the VaR estimates and causing the unusual behavior in the plots.
At this stage, the exact cause is not clear. It likely involves a mix of issues with the historical simulation framework and the dataset. More investigation is needed to fully understand the problem.

# 5 Conclusion and Discussion

This thesis examined the forecasting and testing of Value-at-Risk (VaR), an essential tool for managing financial risks. Using ten years of S&P 500 data, several methods for estimating VaR were applied and evaluated for accuracy through backtesting. The study revealed that traditional methods like Historical Simulation (HS) and Variance-Covariance (VC) performed reasonably well during stable market periods but struggled during volatile times, such as the COVID-19 crisis in 2020. These methods often underestimated risk during extreme events because they assume market conditions remain constant over time.

Conditional methods, such as EWMA and GARCH, proved to be more effective as they adapt to changing market conditions. Among these, EWMA-based methods consistently provided reliable results. The Variance-Covariance approach using a t-distribution (VC_t) addressed the issue of rare, extreme losses (fat tails) but tended to overestimate risk in calmer periods. Backtesting results showed that many methods produced the expected number of breaches, but some, like HS and VC, had clustering issues where breaches were not evenly spread out. In contrast, EWMA-based methods, particularly HS_EWMA and VC_EWMA, excelled in adapting to changing market conditions and delivering dependable estimates. However, the HS_GARCH methods displayed unusual behavior, producing no breaches at all, which suggests potential challenges in integrating GARCH volatility modeling with the historical simulation approach.
The year 2020, marked by high market volatility, further underscored the limitations of static methods like HS and VC, which failed to keep up with rapid changes in risk. In contrast, EWMA-based methods demonstrated their strength by providing accurate estimates even in such a challenging environment. Overall, this study highlights the importance of using adaptive models like EWMA and GARCH for effective risk management. These models offer greater reliability and flexibility in dynamic and uncertain market conditions, though there is always room for further refinement and innovation in risk estimation.

This study offered useful insights, but there are some limitations. The unexpected behavior of HS-GARCH models needs further exploration to understand and fix the issues. Additionally, the analysis was limited to the S&P 500 index, which may not fully apply to portfolios with more complex structures or multiple risk factors. Managing risk is crucial for financial markets. This study showed that methods that adapt to changing conditions, like EWMA and GARCH, are more reliable than simpler, static methods. However, there is always room for improving these models to handle new challenges in the future.

## References


[1] Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques and tools*. Princeton university press, 2005.

[2] Philippe Jorion. *Financial risk manager handbook*. John Wiley & Sons, 4th edition, 2007.

[3] John Hull. *Risk management and financial institutions*. John Wiley & Sons, 4th edition, 2015.

[4] Svetlozar T Rachev, Stoyan V Stoyanov, and Frank J Fabozzi. *A probability metrics approach to financial risk measures*. John Wiley & Sons, 2011.

[5] Paul Pfeiffer. The Quantile Functions. https://stats.libretexts.org/Bookshelves/Probability_Theory/Applied_Probability_%28Pfeiffer%29/10%3A_Functions_of_Random_Variables/10.03%3A_The_Quantile_Function?utm_source=chatgpt.com.

[6] Kyle Siegrist. Distribution and Quantile Functions. https://stats.libretexts.org/Bookshelves/Probability_Theory/Probability_Mathematical_Statistics_and_Stochastic_Processes_(Siegrist)/03%3A_Distributions/3.06%3A_Distribution_and_Quantile_Functions.

[7] Martin Haugh. Quantitative Risk Management Basic Concepts and Techniques of Risk. https://www.columbia.edu/~mh2078/QRM/BasicConceptsMasterSlides.pdf. Accessed: 2024-12-02.

[8] Philippe Jorion. *Value at Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill, 3rd edition, 2007.

[9] Jacob Boudoukh, Matthew Richardson, and Robert Whitelaw. The best of both worlds. *Risk*, 11(5):64–67, 1998.

[10] Kevin Dowd. *Measuring market risk*. John Wiley & Sons, 1st edition, 2002.

[11] Olli Nieppola. Backtesting value-at-risk models. Master's thesis, 2009.

[12] Yuanyuan Zhang and Saralees Nadarajah. A review of backtesting for value at risk. *Communications in Statistics-Theory and methods*, 47(15):3616–3639, 2018.

[13] Paul H Kupiec et al. *Techniques for verifying the accuracy of risk measurement models*, volume 95. Division of Research and Statistics, Division of Monetary Affairs, Federal Reserve Board, 1995.

[14] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

[15] Michele Costa, Giuseppe Cavaliere, and Stefano Iezzi. The role of the normal distribution in financial markets. In *New Developments in Classification and Data Analysis: Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Bologna, September 22–24, 2003*, pages 343–350. Springer, 2005.

[16] Emmanouil Karimalis, Paul Alexander, and Fernando Cerezetti. Testing the test: How reliable are risk model backtesting results. https://bankunderground.co.uk/2016/01/15/testing-the-test-how-reliable-are-risk-model-backtesting-results/.