How Many Stocks Should You Buy?

A Simulation Study on Portfolio Diversification for the Swedish Stock Market

Antonio Prgomet

Master's thesis
2021:E20

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Abstract

For every stock investor, the question of how many stocks to buy is fundamental. The recommendations from the literature is wide and ranges from 10 to over 300. As a contrast, 41.79% of Swedish shareholders held only one stock in year 2020. This thesis studies the Swedish stock market between 2011 and 2020 which to the authors knowledge not has been done before within the research space of Diversification. Main conclusions are that for equally weighted portfolios, reducing the shortfall risk to 20% for a 10 year investment period would take around 120-150 stocks. For five (one) year periods it would take 150 stocks (70-80 stocks). It was also seen that portfolios held for 10 years second-order stochastically dominates portfolios held for five years and one year respectively when the portfolio size is five and greater for equally weighted portfolios.

Keywords Portfolio Diversification, Modern Portfolio Theory, Quantitative Finance, Return Distributions, Shortfall Risk, Stochastic Dominance, Simulation Study.

Acknowledgement

I am fortunate to have been a student to Associate Professor Magnus Wiktorsson. He taught me important things and supervised me in this thesis with commitment.

Professor Erik Lindström introduced me to modern research within Portfolio Theory which is the domain of this thesis.

The many discussions with Jakob Lavröd gave me new perspectives and insights throughout my studies and when writing this thesis.

I would like to thank my friends: Per Niklas Waaler, Jakob Lavröd, Tien Truong, Joel Persson, Marlon Almström, Heidi Mach, Hannes Nilsson, Carl Lokrantz, Henning Tansjö, Thomas Hamfelt, Arsen Kadoian, Jacob Karlsson and Emma Ryborg. Without you it would not be as fun as it was.

I am grateful to my family for their continuous and unconditional support.

Finally I wish to thank the Mathematical Department and Economics & Business Department for providing a good education that have made me well prepared for the interesting challenges awaiting.

Contents

Chapter 1

Introduction

Diversification or the practice of investing one's money into many different assets aims to reduce exposure to one particular asset or risk. The concept is not new, in the old testament, Ecclesiastes 11:2, it is written: *"Divide your investments among many places, for you do not know what risks might lie ahead"*. In common parlance, the same principle is expressed through the phrase *"Don't put all your eggs in one basket"*. Assuming that the stocks do not perform exactly the same, the intuition behind diversification is simple. Buying shares in only one company, there is a risk that you buy the worst performing stock. If you instead buy two stocks, the worst case scenario is that you buy the two worst performing stocks, but still you are better off than only buying one stock since the second worst performer is still better than the worst performer due to our assumption that the stocks do not perform exactly the same. On the other hand, buying one stock you could buy the best performer, leaving you better off compared to buying the two best performers. So having concentrated portfolios will make your portfolio more volatile than spreading your investments and is therefor perceived as more risky. The most straightforward way to reduce this risk is through diversification.

With Harry Markowitzs' groundbreaking work [1952, 1959], a theoretical framework justifying the benefits of diversification in a mathematically rigorous way was established. With it, the investment choice was pushed away from being an art to a science and what today is called the "Modern Portfolio Theory" was born. Capital Asset Pricing Model (CAPM), a model directly based on Markowitzs' work, implies that the only efficient portfolio is the market portfolio consisting of all risky assets each weighted in proportion to its total presence in the market. With CAPM, the existence of Index Funds and extensive diversification can be theoretically justified.

Several empirical studies have been conducted on the question *"How many stocks does it take to be well-diversified?"* with recommendations ranging from 10 to over 300 stocks which most likely is quite surprising results for the reader. Evans and Archer [1968] Suggest around 10 stocks, Elton and Gruber [1977] recommend atleast 15 stocks, Benjelloun [2010] recommends around 40-50 stocks and Statman [2004] recommends over 300 stocks. This wide range, is no doubt troublesome for the investor facing an actual investment decision. How many stocks should he buy to be well-diversified? As time passes, changing market conditions likely do change the number of how many stocks make a diversified portfolio. Similarly, a different number is expected if different markets (such as Large Cap stocks in Sweden or S&P 500 in the US) are used. But most notably, different methodologies to measure the gain of diversification will change the recommendations suggested in a direct manner. In the litterature, different methodologies have been used and is a major reason for the wide range of suggestions.

These studies can be contrasted to the current situation in Sweden. In year 2020 (2019), the average number of stocks held per shareholder is 4.5 (4.1) and 41.79% (43.9%) held only one stock [1] [Euroclear, 2020, p.12].

1.1 Research Questions

The question of "How many stocks does it take to be well-diversified" has a established research tradition (see section 2.6) and is a fundamental question for every investor. To mine knowledge, no study in this research space has looked at the Swedish stock market as this study do. The thesis also looks at "modern time" spanning from 2011-01-03 to 2020-12-30 making it relevant for current investors. The question of whether equally weighted or capitalization weighted portfolios are preferable will also be studied. Finally, whether one, five or 10 year investment periods are better will be studied.

[1] Many people are in practice more diversified by for instance having invested in real estate or through part of their pension savings which are in some funds.

Chapter 2

Theoretical Background and Litterature Review

In this section, theory and literature relevant for understanding the context of this thesis is presented. Especially section 2.6 is important for understanding the surrounding literature of this thesis.

2.1 Overview of Modern Portfolio Theory

In his seminal 1952 paper, Markowitz introduced what today is known as Modern Portfolio Theory and with it, pushed investment choice away from beeing an art to a science. The central message of the theory is simple. Investors should invest in portfolios that for a given target expected return minimizes risk as measured by variance [1] or alternatively maximize portfolio expected return given a target variance. Mathematically, this can be stated as a quadratic optimization problem:

$$\mathbf{w} = \arg\min_{\mathbf{w}} \mathbf{w^T \Sigma w} \tag{2.1}$$

subject to the constraints:

[1]The individual investor can use the risk measure of his choice, for instance semivariance which is defined as: $\mathbb{E}(X - \mathbb{E}X)^2 \mathbb{1}_{\{X \leq \mathbb{E}X\}}$. The standard deviation is also commonly used which is equivalent to the variance (in a optimization context) due to there being a one-to-one correspondence between standard deviation and variance, they are strict monotone transformations of each other.

$$\mathbf{w^T}\boldsymbol{\mu} = \mu^* \tag{2.2}$$

$$\mathbf{w^T}\mathbf{1} = \sum_i w_i = 1 \tag{2.3}$$

where $\mathbf{w}$ is the column vector of relative portfolio weights, $\boldsymbol{\Sigma}$ is the covariance matrix, $\boldsymbol{\mu}$ is the column vector consisting of the expected returns of individual assets and μ^* is the, exogenously set, target expected portfolio return which is a scalar. The final constraint, equation (2.3), means that the portfolio is fully invested.

In this case, the optimization problem, has a closed form solution which can be found by using Lagrangian multipliers, for a complete derivation see Appendix A. The solution is:

$$\mathbf{w_{optimal}} = \frac{C - \mu^* B}{AC - B^2}\boldsymbol{\Sigma}^{-1}\mathbf{1} + \frac{\mu^* A - B}{AC - B^2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$$

with $A = \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}$, $B = \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}$ and $C = \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.

Often, prohibiting short sales is also added as a constraint, i.e. $w_i \geq 0$. In this case the problem has no closed form solution but solving it is still easy by using quadratic programming solvers, such as quadprog in MATLAB.

The well-known CAPM model by Sharpe [1964] is directly based on the work of Markowitz and is constructed as an Equilibrium argument. One key implication of the model is that the only efficient portfolio is the "market portfolio" containing *all* risky assets where each asset is weighted by its market capitalization. So practitioners of CAPM should engage in extensive diversification. An important underlying assumption in the model is that investors have homogenous expectations of the future, i.e. they all have the same expectations regarding the expected value, variance and correlation coefficients of all assets in the investment universe. While this might be a reasonable modelling assumption, in practice, an investor contemplating different investment opportunities has to form an opinion about expected values, variances and correlations of returns himself. Once this is done, it only remains to plug in the estimates in the Markowitz optimization framework to obtain optimal portfolio weights. Despite the conceptual simplicity of this procedure, in practice, results that are bad and not economically meaningful often results [2]. One reason for this is the sensitivity of the optimizer, Michaud coined the term "Error Maximazation" which he describes in the following way:

> *Risk and return estimates are inevitably subject to estimation error. MV optimization significantly overweights (underweights) those securities that have large (small) estimated returns, negative (positive) correlations and small (large) variances. These securities are, of course, the ones most likely to have large estimation errors. [Michaud, 1989, p.33-34]* .

The following example, taken from DemMiguel, Garlappi and Uppal [2009, p.1919-1920] strengthens the intuition and shed further light on the important problem of "Error Maximazation".

Suppose we have two assets where both have the same yearly mean and volatility of 8% and 20% and with a correlation of 0.99. Due to the both assets being identical, the optimal weight allocation from the optimizer is 50% in both assets. But, if the mean return on the first asset is not known and is estimated to be 9% instead of 8%, then the mean-variance optimizer would recommend a weight of 635% in the first asset and -535% in the second. So, as this example demonstrates, the optimizer try to exploit even the smallest difference by extreme weighting without considering that these differences might be due to simple estimation error which is a most realistic assumption in all statistical problems including covariance- and mean estimates.

The same paper by DeMiguel et al. [2009] also demonstrates that it is hard to consistently beat the naive $1/N$ portfolio. This is a very serious problem for those wishing to use the mean-variance analysis for practical allocation problems. Several approaches to deal with practical limitations of the opti-

[2] One interesting fact in the context is that Markowtiz himself used naive diversification rather then his own framework for personal investments. *"I should have computed the historical co-variances of the asset classes and drawn an efficient frontier. Instead, I visualized my grief if the stock market went way up and I wasn't in it–or if it went way down and I was completely in it. My intention was to minimize my future regret. So I split my contributions fifty-fifty between bonds and equities"* [Zweig, 1998, p.114]

mization problem exist and I will provide a short overview.

Jorion [1992] recognizes that the classical optmization alghoritm ignores the effect of estimation in the input estimates. Therefore he proposes a simple simulation approach based on the Multivariate Normal Distribution so the investor can obtain a range of "statistical equivalent portfolios". Another resampling technique, called "The Resampled Efficient Frontier" is a method based on Monte Carlo resampling. See chapter 6 in the book Michaud and Michaud [2008] for details.

To estimate the covariance matrix needed for portfolio optimization, a natural first step is to use its sample equivalent. In situations when there are few observations relative to the number of stocks, as is often the case [3], large estimation errors occur. Hence, the most extreme coefficients in the matrix tend to take extreme values, not so much because this is the "truth" but rather since they contain alot of error. Unfortunately, due to "Error Maximization", the optimizer will tend to focus on those coefficients that are most unreliable. To remedy this, Ledoit and Wolf [2003] and Ledoit and Wolf [2004] suggest the method of linear shrinkage where extreme coefficients are pulled towards more central values using a shrinkage target and thereby reducing estimation error where it matters the most. Mathematically it can be formulated as:

$$\Sigma_{shrink} = \delta F + (1 - \delta)S$$

where Σ_{shrink} is the shrunken covariance matrix, F is the target, S is the sample covariance matrix and δ is the optimal shrinkage intensity. Different targets can be used, Ledoit and Wolf [2003] use for instance the single factor matrix due to Sharpe [1963] and Ledoit and Wolf [2004] use a constant correlation model where all pairwise correlations are identical. One crucial part that remains is how to choose the optimal shrinkage constant δ. Ledoit and Wolf [2003, 2004] derives the optimal shrinkage intensity using a loss function based on the Frobenius norm. For nonlinear shrinkage, see for instance Ledoit and Wolf [2017] which they have shown dominates linear shrinkage. For a paper combining nonlinear-shrinkage with multivariate GARCH models, see Engle, Ledoit and Wolf [2009]. In that paper they consistently beat

[3]For example when the number of stocks considered is large and perhaps five years of monthly returns are used yielding 60 observations.

the $1/N$ portfolio which DeMiguel et al. [2009] shows is a non trivial achievement.

For mean estimates, the reader can for instance start by reading the paper by Green et al. [2013] which gives a broad overview.

One approach starting the asset allocation based on an equilibrium argument and allowing the investor to incorporate his own opinion, is the Black-Litterman model. See for instance Black and Litterman [1992].

2.2 Stochastic Dominance

In this section, the concept of second-order stochastic dominance will be introduced. It is a partial order[4] which gives an explicit and transparent rule for ordering portfolios and random variables in general. The concept of second-order stochastic dominance is what will be used in this thesis but to understand it better, first-order stochastic dominance is introduced as well.

There is a rich theory related stochastic dominance and here only the essentials are presented. The interested reader can for instance look at the book: Whang [2019] and articles: Levy [1979] and Domian et al. [2007] as a start for further studies.

We will focus on continuous distributions here despite that the concept of Stochastic Dominance can be defined for discrete or mixed distributions as well.

We begin by defining first-order stochastic dominance.

Definition 1. *The random variable X_1 first-order stochastically dominate the random variable X_2, if $P(X_1 \geq x) \geq P(X_2 \geq x)$ for all x and $P(X_1 > x) > P(X_2 > x)$ for some x.*

[4]It is a partial order since for example, if portfolio A does not second-order stochastically dominate portfolio B, that does not imply that Portfolio B second-order stochastically dominates portfolio A. As a comparison, for a total order all pairs are comparable.

□

Remark 2. *Visually, X_1 first-order stochastically dominates X_2 if the cumulative distribution function (CDF) of X_1 lies below that of X_2.*

□

Example 3. *Suppose the randomo variables X_1 and X_2 denote the return of two stock portfolios. If X_1 first-order stochastically dominates X_2, then the first portfolio would be desired by all investors preferring higher returns.*

□

We continue to the main concept of second-order stochastic dominance.

Definition 4. *The random variable X_1 second-order stochastically dominate the random variable X_2, if for all $x \in \mathbb{R}$*

$$\int_{-\infty}^{x} [F_1(t) - F_2(t)]\, dt \geq 0$$

and there is strict inequality for some x, i.e.

$$\int_{-\infty}^{x} [F_1(t) - F_2(t)]\, dt > 0$$

□

Proposition 5. *First-order stochastic dominance implies second order-stochastic dominance.*

Proof. Follows immediately from definition 1 and definiton 4. □

Theorem 6. *Assuming an investor to be a risk averse (concave utility function) and expected-utility maximizer, he would prefer a portfolio X_1 if it second-order stochastically dominates portfolio X_2.*

Proof. Proof is omitted. □

Remark 7. *The reader should notice that the theorem is a very strong and compelling argument for ranking different portfolios given that the investor is a risk averse expected-utility maximizer.*

2.3 Some Mathematical Preliminareis

We will provide some mathematical formulae that are important for understanding Diversification and Portfolio Theory in general. Although important they are not that advanced. For completeness proofs are provided.

Lemma 8. *A double sum can be written as a product of two sums:*

$$\sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j = \sum_{i=1}^{n} x_i \sum_{j=1}^{m} y_j \tag{2.4}$$

Proof.

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j &= (x_1 y_1 + ... + x_1 y_m) + \ldots + (x_n y_1 + ... + x_n y_m) \\
&= (x_1 + ... + x_n) y_1 + ... + (x_1 + ... + x_n) y_m \\
&= \left(\sum_{i=1}^{n} x_i\right)(y_1 + ... + y_m) \\
&= \sum_{i=1}^{n} x_i \sum_{j=1}^{m} y_j
\end{aligned}$$

□

Definition 9. *We define the covariance, variance and standard deviation.*

$$Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}Y)],$$

$$Var(X) = \mathbb{E}[((X - \mathbb{E}(X))^2],$$

$$Std(X) = \sqrt{Var(X)}.$$

□

From the definition, it is immediate that the variance can be expressed in terms of the covariance, i.e. $Var(X) = Cov(X,X)$.

Lemma 10. *$Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.*

Proof. The proof follows immediately by multiplying out the brackets in the definition of covariance, and then using the linearity of mathematical expectation.

$$\begin{aligned} Cov(X, Y) &= \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}Y)] \\ &= \mathbb{E}[XY - X\mathbb{E}(Y) - \mathbb{E}(X)Y + \mathbb{E}(X)\mathbb{E}(Y)] \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \end{aligned}$$

□

Proposition 11. *Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be random variables. Let $a_1, \ldots, a_m$ and $b_1, \ldots, b_m$ be real numbers. Then*

$$Cov\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j).$$

Proof. Using lemma 8 that a double sum is the product of two sums, lemma 10 and linearity of expectation, we obtain the following:

$$\begin{aligned} Cov\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_i\right) &= \mathbb{E}\left[\sum_{i=1}^{n} a_i X_i \sum_{j=1}^{m} b_j Y_j\right] - \mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] \mathbb{E}\left[\sum_{j=1}^{m} b_j X_j\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \mathbb{E}(X_i, Y_j) - \sum_{i=1}^{n} a_i \mathbb{E}\left(X_i\right) \sum_{j=1}^{m} b_j \mathbb{E}\left(Y_J\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \mathbb{E}(X_i, Y_j) - \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \mathbb{E}(X_i)\mathbb{E}(Y_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \left[\mathbb{E}(X_i, Y_j) - \mathbb{E}(X_i)\mathbb{E}(Y_j)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j). \end{aligned}$$

□

Since $Var(X) = Cov(X, X)$, we get the following corollary, which is a general expression for the variance of a linear combination.

Corollary 12.

$$Var(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j Cov(X_i, X_j).$$

Proof. From the equality $Var(X) = Cov(X, X)$ and proposition 11, we get the chain of equalities:

$$Var(\sum_{i=1}^{n} a_i X_i) = Cov\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{n} a_j X_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j Cov(X_i, X_j).$$

□

Remark 13. *Identifying terms with the same index, we can split the last double sum in the following way which emphasize the pure variance and covariance contributions respectively.*

$$Var(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j Cov(X_i, X_j) \tag{2.5}$$

$$= \sum_{i=1}^{n} a_i^2 Var(X_i) + \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} a_i a_j Cov(X_i, X_j). \tag{2.6}$$

□

2.4 Sub-Additivity: Justifying Diversification

In Modern Portfolio Theory, diversification is a core concept. Its importance stems from the fact that the investor should not only focus on returns, but on risk aswell. Markowitz writes the following about the importance of Diversification: *"Diversification is both observed and sensible; a rule of behavior* *which does not imply the superiority of diversification must be rejected both as a hypothesis and as a maxim"* [Markowitz, 1952, p.77].

I begin by providing a simple, motivating, example showing the power of diversification.

Example 14. *Suppose stock A has an expected return of 4% and a standard deviation (volatility) of 12%. Assume that Stock B also has a expected return of 4% and a standard deviation (volatility) of 12%. The correlation between the two assets is 0.8. In this example, we assume that $w_A = w_b = 0.5$, i.e. the portfolio is equally invested in the two stocks. We model both stock returns as random variables. With obvious notation, we have for instance $\mathbb{E}(X_A) = 4\%$ and $Std(X_A) = 12\%$. Are we better of buying one of the stocks or a combination of them?*

Solution. *If we buy one of the stocks, we obviously get a return of 4% with a volatility of 12%. But assuming we buy both of the stocks, then the following calculations gives the expected portfolio return ($E[R_p]$) and standard deviation (σ_p) of the portfolio:*

$$E[R_p] = E\left[(0.5X_A) + 0.5X_B\right] = 4$$

In words, we get an 4% expected return, which is unchanged compared to buying one stock.

$$Var\left[(0.5X_A) + 0.5X_B\right] = 0.5^2Var(X_A) + 0.5^2Var(X_B) + 2(0.5^2)Cov(X_A, X_B)$$

Since Correlation, $Corr(X,Y)$ is defined as $Corr(X,Y) = \frac{Cov(X,Y)}{Std(X)Std(Y)}$, a simple calculation shows that $Cov(X_A, X_B) = 0.01152$. Using this and remark 13, we get that

$$Var\left[(0.5X_A) + 0.5X_B\right] = 0.5^2 0.12^2 + 0.5^2 0.12^2 + 2(0.5^2)0.01152 = 0.01296$$

And finally, we obtain, $\sigma_p = \sqrt{0.01296} \approx 0.1138$.

□

This example shows the gain of diversification. By buying two stocks, we get to keep our return of 4% but lower our standard deviation from 12% to approximately 11.38%. Obviously, the question is whether we can expect this in the general case? The answer is yes, and it is follows from the sub-additivity property of the standard deviation.

In a two asset case, we can denote the return distributions as random variables, X and Y. Then the sub-additivity of standard deviation is simple to show. $Var(X+Y) = Var(X) + Var(Y) + 2Corr(X,Y)Std(X)Std(Y)$. Since $Corr(X,Y) \leq 1$, we have that: $Var(X+Y) \leq Var(X) + Var(Y) + 2Std(X)Std(Y) = [Std(X) + Std(Y)]^2$, taking square root on both sides, we finally get: $Std(X+Y) \leq Std(X) + Std(Y)$. In words, this says that if adding two assets together, the risk/volatility (as measured by standard deviation) cannot get bigger than if you add the two risks seperately. So you are never worse off, in terms of volatility, if you divide your money into more assets. Example 14 illustrated this in a special case. To stress the interpretation of sub-additivity, I quote Willmotts explanation of it: *"If you add two portfolios together the total risk can't get any worse than adding the two risks seperately. Indeed, there may be cancellation effects or economies of scale that will make the risk better"* [Wilmott, 2006, p.342].

We now prove sub-additivity for standard deviation in the general case when different weights are used for different assets and where we have more than two assets.

Proposition 15. *Denote by $w_i \geq 0$ the weight in asset i. We model returns as random variables X_i, where the index i represents asset i. Then, the portfolio standard deviation is less than or equal to, a linear combination of the return standard deviations of the portfolio constituents. The weights in the linear combination are the portfolio weights. Mathematically we have:*

$$Std(\sum_{i=1}^{n} w_i X_i) \leq \sum_{i=1}^{n} w_i Std(X_i)$$

Proof. We first work with variances and will at the end take the square root to obtain the result for standard deviations.

$$\begin{aligned} Var(\sum_{i=1}^{n} w_i X_i) &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j Cov(X_i, X_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j Std(X_i) Std(X_j) Corr(X_i, X_j) \end{aligned}$$

Where we used that $Cov(X_i, X_j) = Corr(X_i, X_j)Std(X_i)Std(X_j)$. Since $Corr(X_i, X_j) \leq 1$, we get:

$$Var(\sum_{i=1}^{n} w_i X_i) \leq \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j Std(X_i) Std(X_j) = \left(\sum_{i=1}^{n} w_i Std(X_i)\right)^2$$

Taking square root on both sides, we finally get:

$$Std(\sum_{i=1}^{n} w_i X_i) \leq (\sum_{i=1}^{n} w_i Std(X_i))$$

This shows that the portfolio standard deviation is less than or equal to, a linear combination of the return standard deviations of the portfolio constituents, where the weights in the linear combination are the portfolio weights. So the risk of a portfolio is less than, or equal to, the risk of the individual assets. □

When solving the optimization problem to obtain optimal weights it does not matter whether the investor uses standard deviation or variance. This since there is a one-to-one correspondence by virtue of the two measures being a strict monotone transformation of each other. But, in terms of evaluating risk in itself, there is a big difference. Whereas standard deviation is sub-additive which implies diversification never leaves you worse off, variance is not. A simple counterexample proves this.

$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$, assuming $Cov(X,Y) > 0$, implies that $Var(X+Y) > Var(X) + Var(Y)$. If we let $X = w_1X_1$ and $Y = w_2X_2$, we have translated the above proof to the notation used above in a portfolio theory context.

In general, it is possible to show that standard deviation is a coherent risk measure whereas variance is not, hence the two measures are not equivalent in general.

Markowtiz writes *"Thus far I have used the standard deviation (or equivalently, the variance) of return as a measure of the risk involved in the portfolio"* [Markowitz, 1976, P.50]. Due to the discussion above, these kind of statements are in general not correct and care must be taken regarding the context.

2.5 The Law of Average Covariance and Diversifiable Risk

With the Mathematical machinery set, we can readily obtain an analytical formula to evaluate the benefits of naive diversification. It will also equip us with an way to decompose total risk into diversifiable and non-diversifiable risk.

Recall from equation (2.6), that the portfolio variance, denoted by σ_p^2 here, can be written as:

$$\sigma_p^2 = \sum_{i=1}^{n} w_i^2 Var(X_i) + \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} w_i w_j Cov(X_i, X_j) \tag{2.7}$$

If we naively, choose each portfolio weight as $1/N$, where N is the number of assets, then equation (2.7) becomes:

$$\sigma_p^2 = \left(\frac{1}{N}\right)^2 \sum_{i=1}^{n} Var(X_i) + \left(\frac{1}{N}\right)^2 \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} Cov(X_i, X_j). \tag{2.8}$$

In an N asset case, there are N variances, and $N^2 - N = N(N-1)$ covariances. Naturally then, we can define the average variance and average covariance respectively as:

$$\overline{\sigma_p^2} = \frac{1}{N}\sum_{i=1}^{n} Var(X_i)$$
$$\overline{\sigma_{ij}} = \frac{1}{N(N-1)}\sum_{i=1}^{n} Cov(X_i, X_j)$$

Using the fact that $\frac{1}{N^2} = \frac{N-1}{N}\frac{1}{N(N-1)}$, we can rewrite equation (2.8) as:

$$\sigma_p^2 = \frac{1}{N}\overline{\sigma_p^2} + \frac{N-1}{N}\overline{\sigma_{ij}} \tag{2.9}$$

Taking the limit as N goes to infinity in equation (2.9) we see that:

$$\sigma_p^2 \to \overline{\sigma_{ij}}. \tag{2.10}$$

In words, the risk that remains after diversifying naivly among infinitely many stocks is the average covariance. Hence, this is the risk that we cannot diversify away. In the special case that the average covariance is 0, we could eliminate risk completely. While empirically unreasonable for stocks, this might be a more realistic situation for insurance companies for instance, assuming insurance policies are written on uncorrelated risks.

Note that in this theoretical example, covariances and not variances are the important factor. In other words, evaluating securities in themselves, ignoring the covariance between them, is a bad idea for the portfolio manager. He should look at the portfolio as a whole and not only as consisting of individual parts. Also note that in general, for N securities, there are N

variances but $N(N-1)$ covariances. Assuming there are 200 assets in the investment universe, only 200 variance estimates are needed compared to the $200 \times 199 = 39\,800$ covariance estimates that are needed. This explains why covariances often are more important than variances when creating portfolios and very soon can become an overwhelming task if too many securities are analyzed. This "curse of dimensionality" problem can for instance be handled by the single index model of Sharpe [1963], see the article for details.

What we derived here was by Markowitz called "The Law Of The Average Covariance" [1976]. It was however derived much before in his classical book

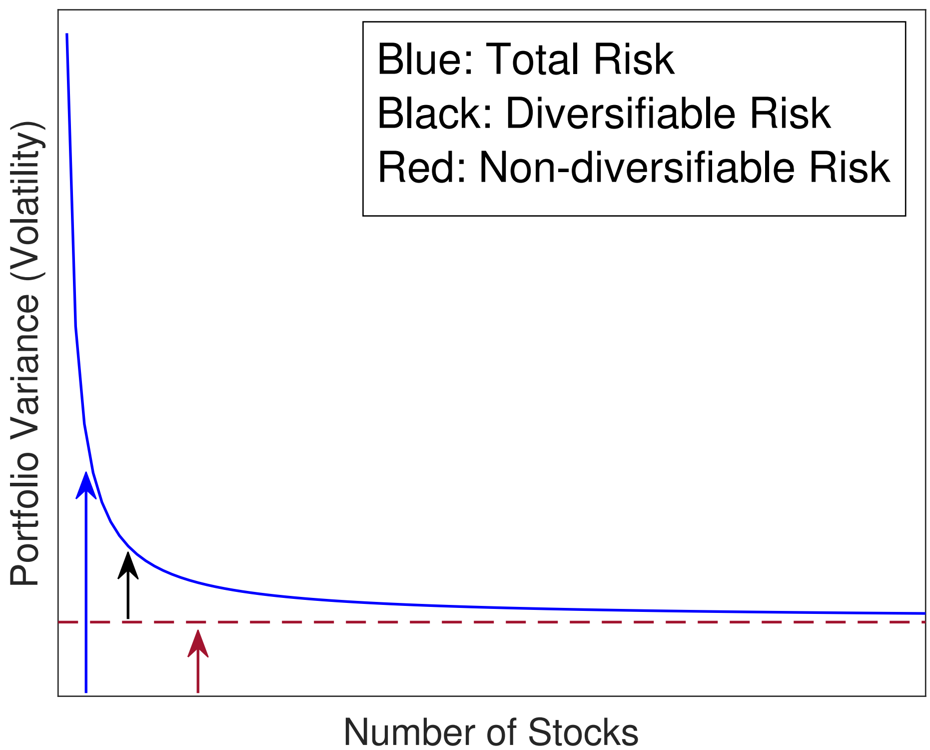

Figure 1: Total Risk can be decomposed as: Total Risk = Diversifiable Risk + Non-diversifable Risk. Blue arrow shows total risk, black arrow shows diversifiable risk and red arrow shows non-diversifiable risk.

[1959, see p.109-112] but was then presented without a name.

In this section we obtained the "Law of Average Covariance" by virtue of the simplifying assumptions of equal weighting and infinitely many assets. In reality this might be non-realistic assumptions. Nevertheless they are necessary in order to obtain nice close form formulas as in this section which equips the investor with intuition and better understanding of the concept of diversification.

Having obtained a mathematical working knowledge of diversification and related formulas, in next section we continue to describe some empirical work that has been done in the topic of diversification and is very important for understanding the context of this thesis.

2.6 How Many Stocks Does It Take To Be Well Diversified?

Modern Portfolio Theory recognizes the gain of diversification. Naturally, the questions arises, how many stocks make a diversified portfolio? Several studies has attempted to answer this question with different data sets and different methodologies. While changing market conditions, naturally, can affect the number of stocks needed to be well diversified, it would be very troublesome if that number ranged from say 10 to 300 stocks due to changing market conditions. Fortunately, this wide range of suggestions can by a large extent be explained by the fact that fundamentally different methodologies have been used in answering the question "how many stocks does it take to be well diversified".

The aim of this section is to give a broad overview of some studies and the reader who wish details is encouraged to read the original articles which are very much readable. A summary of the number of stocks suggested is available in table 1 at (p.23).

The pioneering study attempting to measure the relationship between increasing portfolio diversification and reduction of portfolio risk is attributed to Evans and Archer [1968]. Their methodology is conceptually simple. They built portfolios based on random selection with equal weighting of stocks from a universe consisting of 470 [5] stocks from the Standard and Poor's Index (S&P 500) and calculated the time-series standard deviation of their semi-annual return between January 1958 - July 1967. The portfolio sizes ranged form $1-40$. They repeated this 60 times to obtain 60 observations of the standard deviation for each portfolio size. With 60 observations at hand, they calculated the mean portfolio standard deviation for each portfolio size and plotted this against the portfolio size. Figure 1 (p.17) catches the spirit of the plot they obtained. They also ran a regression [6] $Y = B(1/X) + A$, where Y is the mean portfolio standard deviation, X is the portfolio size and A is the constant. From the regression, they concluded that the mean standard deviation decreased to an asymptote and that this asymptote approximated the average estimated systematic variation [7] of 0.1166 over the period considered. they conclude that the results raise doubts concerning the economic justification of icreasing portfolio sizes beyond 10 or so securities.

The study Evans and Archer did had a large impact and was used as a theoretical motivation when suggesting how many stocks an investor should buy in many textbooks for example. See Newbould and Poon [p,73, 1993] and [p.85, 1996] respectively for a survey of textbooks and more popular publications.

Newbould and Poon [1993] recognizes that several authors have focused on the *average* standard deviation of portfolios with different sizes, while an investor only has one portfolio. Hence, the investor faces the risk that his specific portfolio can substantially be below or above the average. As they themselves state; *"Each individual investor is risk averse on his/her own individual portfolio outcome (and not merely risk averse on the average of all investors' outcomes)"* [1993, p.87]. They remedy this by simply constructing confidence intervals. They conclude that the minimum number of stocks needed to achieve diversification is much higher than 20 stocks but that it

[5] 30 stocks were removed due to unsatisfactory data.

[6] Bird and Tippett [1986] shows that their regression specification in general is misspecified. The consequence is biased parameter estimates which, in this specific case, translates to over-estimation of the rate at which diversifiable risk is eliminated as the portfolio size is increased due to an under-estimate of B.

[7] Calculated by computing the standard deviation of a portfolio containing all the 470 securities.

ultimately depends on the particular stock universe studied, the weighting scheme used when constructing the portfolios and the individual investors preferences. It should be noted that Evans and Archer [1968] also construct a confidence interval but do not emphasize its importance in their study.

Upson et al. [1975] write that another aspect of risk in context of diversification is a portfolio managers confidence of being near the average market return, i.e. the variability across portfolios for a given time period is relevant. So there is a risk in the cross-sectional direction as well in the sense that portfolio returns for specific portfolio sizes, can vary more or less from the market portfolio. Considering this with an example, they conclude that the "established truth" (at their time) of $8-16$ stocks being enough for diversification is misleading and too low. Elton and Gruber [1977] derives an analytical risk measure incorporating both the time-series dimension as-well as the cross-sectional dimension (portfolio return differing from the market return) and call this "Total Risk". Their conclusion from an empirical study is that the gains from adding stocks beyond 15 appears to be significant. Newbould and Poon [1996] writes that *"Volatility in performance must be "felt" more keenly by the investor than the volatility of standard deviations"* [p.75]. They remedy this by constructing a method where choice of portfolio size (i.e. how much to diversify), is dependent on confidence intervals for average risk and confidence intervals for average returns. An investor investing in large stocks (minimum stock market value of \$ 2.5 billions) wanting to be within 10% of average risk and 10% of average return (with 95% confidence) would need at least a portfolio consisting of 60 stocks. The same scenario but looking at small stocks (stock market value between \$ 150 million and \$ 1 billion) would require at least 100 stocks.

Benjelloun [2010] studies diversification by looking at equally weighted and market weighted portfolios. He studies two measures of risk, time series standard deviation and standard deviation of terminal wealth which is cross-sectional. Benjelloun run regressions of the form $Y = A\frac{1}{N^2} + B$ where N is the portolio size and Y is the measure of risk. He declares a portfolio "diversified" when its risk is equal to or smaller than B. He concludes that 40 to 50 stocks is all that is needed to be diversified.

Mao [1970] does an theoretical analysis and assumes that all securities have the same mean return, the same variance of return and that all pairs of securities have the same positive correlation coefficient. He then derives a formula where he divides the actual gain of diversification by the maximal gain of diversification and calls this the relative gains of diversification. Setting the correlation coefficient (which is the same for all pairs of stocks by assumption) to 0.5, then the number of stocks needed varies between 2.7 to 16.3 to achieve a relative gain between 50% and 90%. If the correlation coefficient is set to 0.2, then the number of stocks needed varies between 4.4 to 33.2 to achieve a relative gain between 50% and 90%. He concludes that relatively few securities are needed to reap the bulk benefits of diversification.

Markowitz [1976] shows by an argument using the "Law of Average Covariance" (see section 2.5), and assuming pairwise correlation of 0.25, that much of the reduction in standard deviation that diversification will yield is already provided by 20 stocks.

Evans and Archer [1968] recognize the importance of an marginal analysis when deciding what portfolio size to have which is a fundamental, well-known, principle in Microeconomics. Statman [1987] writes that *"Diversification should be increased as long as the marginal benefits exceed the marginal costs. The benefits of diversification are in risk reduction. The costs are tranaction costs."* [p.354]. Using this kind of argument in combination with a security market line which builds on the ability to borrow or lend money (see the paper for details), he concludes that the borrowing investor needs at the very least 30 stocks and the lending investor 40 stocks.Statman [2004] uses the same methodology as in the previous paper, but due to changing market conditions he updates his assumptions. The conclusion is that the optimal level of diversification is over 300 stocks. For another paper using, among other things, the marginal analysis method, see Rui Ming Daryl and Kai Jie Shawn [2012].

Fisher and Lorie [1970] studies gain of diversification by looking at the distribution of returns. These distributions were obtained by selecting a portfolio of different sizes and repeating the procedure with a computer and was hence a simulation study. They also used different measures for measuring diversification. They conclude that roughly 40% of achievable reduction is obtained by holding two stocks, 80% by holding eight stocks, 90% by holding sixteen stocks, 95% by holding thirty-two stocks and finally 99% by holding 128 stocks. Domian et al. [2003] also do a simulation study and define shortfall risk as the difference between the return of the stock universe considered and the 5^{th} percentile of the distribution for a given portfolio size, this difference is then divided by the return of the stock universe considered. Notice that the return for the stock universe considered is constant since in each simulation all stocks are selected meaning that there is no variation. If a portfolio is deemed diversified when the shortfall risk is below 10%, then over 60 stocks is needed for 20 year periods and over 40 stocks for 5 year periods.

The two studies above look at the return distributions for different portfolio sizes obtained by simulations and then define measures to measure the gain of diversification. Levy [1979] argues that the whole distribution is relevant to look at and does so by using the concept of second-order stochastic dominance (see section 2.2) . In general, all risk averse expected utility maximizers prefer a second-order stochastic dominant investment more than a dominated one. The result of the study, using the data from the Fisher and Lorie [1970] study, is that an investor prefer 128 stocks (rather than 8, 16, 32 stocks) for 1 year periods, the same goes for 5 year periods. For 10 and 20 year periods, no clear and general results are available and the efficient set consists of 2, 16, 32 and 128 stocks for the 10 year period. For the 20 year period, 8 stocks is always preferred but diversification beyond this may not pay.

Domian et al. [2007] also incorporates, among other things, the second order stochastic dominance methodology. They also look at the shortfall risk of ending wealth beeing less than a target amount, which in their study of 20 years they choose too be the long-term US-treasury bond rate. Their study shows that 63 stocks are needed to achieve a 10% shortfall risk, 93 stocks are needed to make the shortfall risk 5% and 164 stocks are needed to have a 1% chance of under performing the Treasury bonds. They state that the results make a compelling case that 100 stocks are not enough to provide sufficient protection from downside risk.

In Sweden, Aktiespararna [2021] is an independent organization for people saving in stocks. Their recommendation of a well-diversified portfolio is 10-15 stocks in different industries.

Table 1: Summary of selected articles from the littrature regarding: "How many stocks does it take to be well diversified?". The number of stocks recommended is interpreted by the author when no explicit number is specified.

Authors	**Number of Stocks**
Evans and Archer [1968]	less than 10
Mao [1970]	When pairwise correlation coefficient is 0.2: 5 stocks (50% of relative gain), 34 stocks (90% of relative gain)
Fisher and Lorie [1970]	8 (80% risk reduction), 16 (90% risk reduction)
Upson et al. [1975]	More than 16
Markowitz [1976]	Around 20
Elton and Gruber [1977]	Atleast 15
Levy [1979]	128 for 1 to 5-year holding periods, 8 for 20 year holding periods
Statman [1987]	Atleast 30 (borrowing investor), Atleast 40 (lending investor)
Newbould and Poon [1993]	Much higher than 20
Newbould and Poon [1996]	60 (large stocks), 100 (small stocks)
Domian et al. [2003]	Over 40 (5 year periods), over 60 (20 year periods)
Statman [2004]	Over 300
Domian et al. [2007]	Over 100
Benjelloun [2010]	40 to 50
Aktiespararna [2021]	10-15 stocks in different industries

Chapter 3

Method

3.1 Data Collection

Daily closing prices, adjusted for splits, was retrieved from Nasdaq Nordics website [http://www.nasdaqomxnordic.com/aktier/historiskakurser]. Looking at all listed companies from Small-, Mid-, and Large Cap at the date 2021-04-01, companies that had a complete price history from the three lists ranging from 2010-01-04 to 2021-01-04 was then selected.

The data used in this study ranged from 2011-01-03 to 2020-12-30 yielding 10 years of data. I used only this interval of the data (and not all the way back to 2010-01-04) since this gave 10 years of data. This was desirable for consistency reasons, since I could then form 10 one year periods, two five year periods and one 10 year period making the comparisons done in the chapter 4 over the same period. Due to requiring a complete price history back to 2010-01-04 I know that at least one stock was excluded due to this despite that it could have been included if the requirement was that its' price history should range back to 2011-01-03 only. With this, I added an unnecessary bias to the data collection. The seriousness of this bias is not that great and does not invalidate the results of this study.

Market Capitalization values at close was available for the date 2021-03-31 for all stocks included in the study. By dividing the market capitalization value by the closing price for the date 2021-03-31, the number of stocks for the specific company could be deduced. This number was then used throughout the data set to infer market capitalization values for different dates. For instance, multiplying the number of stocks by the closing price 2010-01-04, we obtained a market capitalization value for that date as well. Notice that the assumption of constant number of stocks outstanding is an assumption since companies can for instance repurchase their stocks. Hence, the market capitalization values that was calculated is an proxy for the true market capitalization values.

In total, there were 385 stocks available at 2021-01-04 of which 212 was selected since they had a price history available ranging back to 2010-01-04. One stock, Sagax Preferensaktie, was removed since no market capitalization value was available at the website.

In numbers, this meant that:

- Small Cap had 100 stocks listed 2021-01-04 of which 54 was retrieved.
- Mid Cap had 149 stocks listed 2021-01-04 of which 59 was retrieved.
- Large Cap had 136 stocks listed 2021-01-04 of which 99 was retrieved.

The data does not incorporate dividends, which means that the calculated returns used in the study is an underestimate of the actual returns an investor would obtain in reality. The data collected suffers from survival bias in the sense that only stocks existing at 2021-04-01 was used when collecting data back to 2010-01-04. Hence, companies that for instance went bankrupt 2011 or any other date between the start and end date, was not considered. This kind of bias is not uncommon in the finance literature, see for instance Engle, Ledoit and Wolf [2019, p.370].

That no dividends are incorporated, means that the results of this thesis is pessimistic and the investor would be better off in reality due to the gains from dividends. The survival bias yields more optimistic results since in reality, one cannot see into the future and only selecting stocks that will not go bankrupt. These kind of optimistic results are more problematic since it is in general better to give too careful recommendations than the other way around.

3.2 Study Design - A simulation Approach

The methodology of this thesis is inspired by the methodology presented in the papers Domian et al. [2003] and Fisher and Lorie [1970]. Just as those studies, this study is designed as a simulation study.

Given a return history, the portfolio sizes studied were 1, 2, 5, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 100, 120, 150, 200 and 212 stocks. These portfolio sizes were drawn from a stock universe consisting of 212 stocks from small-, mid-, and large cap [1]. For each of those portfolio sizes, 200.000 simulations were done where the stocks were selected randomly without replacement, i.e. no stock could be chosen two times or more for a specific simulation. The reason why I resort to simulations is due to the number of combinations quickly becoming too big to handle. For example, selecting 1 stock among 212 is possible to do in 212 unique ways which with modern computers is trivial. Selecting 2 stocks among 212, is possible in $\binom{212}{2} = \frac{212!}{(212-2)!2!} = 22.366$ ways and selecting 20 stocks among 212 is possible in $\binom{212}{20} \approx 5.48 \times 10^{27}$ ways. The number of combinations quickly become too large to handle efficiently. Due to the design of the study, we use one return history and simulate different portfolios for each portfolio size. This means that we obtain the distribution of returns for different portfolio sizes *conditional* on the return history used. So, for instance figure 2 (p.31) shows a conditional distribution.

Once the simulations have been done, the percentile thresholds looked at were: 1%, 5%, 10%, 25%, 50%, 75%, 90%, 95% and 99%. Looking for instance at table 3, we see that for portfolios of size 5, there was a 1% that the yearly return was below -0.0039 and it was 50% chance that the return was below 0.1142. Note that the 50$^{\text{th}}$ percentile is the median. Notice, for the stock universe (212 stocks in this case), there is no variation in the percentiles since selecting 212 stocks among 212 can only be done in one way.

Alongside the percentiles, some summary statistics are also presented. These are the mean portfolio return of the 200.000 simulations, the cross sectional standard deviation of the different portfolios and the range defined as the difference between the maximum value and minimum value. Finally, the shortfall in %, is for each portfolio size defined as:

[1] One exception is figure 13 on (p. 64) where only large cap stocks were used.

$$\frac{\text{(Return for the studied stock universe)} - \text{(5\% percentile of a specific portfolio size)}}{\text{(Return for the studied stock universe)}}$$

The choice of using the 5% percentile is not and should not be seen as a "golden standard". I use it here since the 5% value is used in the paper Domian et al. [2003], they on the other hand use the 5% level since it is the customary choice in tests of statistical significance. Nevertheless, each investor has to decide about a suitable level for himself.

[Upson et al., 1975, p.86] argue that that another aspect of risk in context of diversification, is a portfolio managers confidence of being near the average market return. In the same way, [Elton and Gruber, 1977, p.416] claim that earlier studies (such as that one by Evans and Archer [1968]) neglected the risk that the mean return on a specific portfolio will be different than the market return. The shortfall in %, is a response to this kind of critique and was introduced in the paper by Domian et al. [2003].

Periods of one year, five year and 10 years were studied. For ease of interpretation, the returns for five and 10 year periods were annualized making it easy to compare with the one year periods. Only buy and hold strategies were done with no portfolio re-balancing between the start and end date. Equal- and market capitalization weighted portfolios were studied. See section 3.1 for the approximation done regarding the capitalization weighting.

The tables presented in chapter 4 have been averaged. For instance, table 3 is based on one year periods. In total, there were 10 of those available, and the table is the average of these ten observations. Similary, the table presented for five years is the average of two observations. Finally, the ten year table only contained one observation and is not averaged. This way of aggregating the tables provides a better overview and makes it easier to see some general patterns. Nevertheless, care must be taken to not over-interpret the results and for instance think that they are always true for all time periods. Some interesting observations occurring for the specific years are presented in Appendix B.

The programming was done in MATLAB and all results in the thesis are completely reproducible.

Chapter 4

Results and Analysis

I begin by some general comments that are relevant for all periods studied. Next, results that are specific for each period will be commented in the corresponding section of this chapter. This chapter provides an overview, the reader is refereed to Appendix B (p.55) which contains results for specific observations (for instance one year results for 2015-2016 that are not averaged), that deviate from the patterns observed in this chapter.

Looking at figure 2, figure 3 and figure 4, it is clear by a visual inspection that for portfolios of size 5 and greater, the equally weighted portfolios stochastically dominates the capitalization weighted portfolios. Due to this, I shall mainly focus on the tables and figures containing information for the equally weighted portfolios throughout this thesis.

Table 2 summaries the number of stocks needed to reach different levels of shortfall in % for equally weighted portfolios.

Table 2: The number of stocks needed for different periods to reach a specified shortfall in % for equally weighted portfolios.

	Length of period		
Shortfall in %	**1 year**	**5 year**	**10 year**
50	Between 10-20 stocks.	Between 20-30 stocks.	Between 30-40 stocks.
20	Between 70-80 stocks.	Around 150 stocks.	Between 120-150 stocks.
10	Around 150 stocks.	More than 200 stocks.	Almost 200 stocks.

For all tables in this chapter, the cross-sectional standard deviation decreased monotonically with increasing portfolio sizes. For equally weighted portfolios, the mean was not changing much with different portfolio sizes. Doing formal ANOVA tests to test for equal means across all portfolio sizes, we could not reject the null of equal means for one and five year holding periods. For 10 year holding periods, we rejected the null, but looking at table 7 we see that it does not vary too much, in fact the smallest value is 0.2598 (for portfolios of size one) and the largest is 0.2622 (for portfolios of size five).

4.1 One Year Periods

For the equally weighted portfolio, I performed Levene's test for testing the null hypothesis of equal variances across all portfolio sizes in table 3. The null is rejected due to a p-value of 0. Doing the ANOVA test for testing the null hypothesis of equal means across all portfolio sizes in table 3 yielded a p-value of 1, so the null cannot be rejected. These results are as expected when looking at table 3.

Doing the same analysis for the capitalization weighted portfolio yielded a p-value of 0 for Levene's test and a p-value of 0 for the ANOVA test. Again, this is the expected results when looking at table 4, here the mean does vary quite alot.

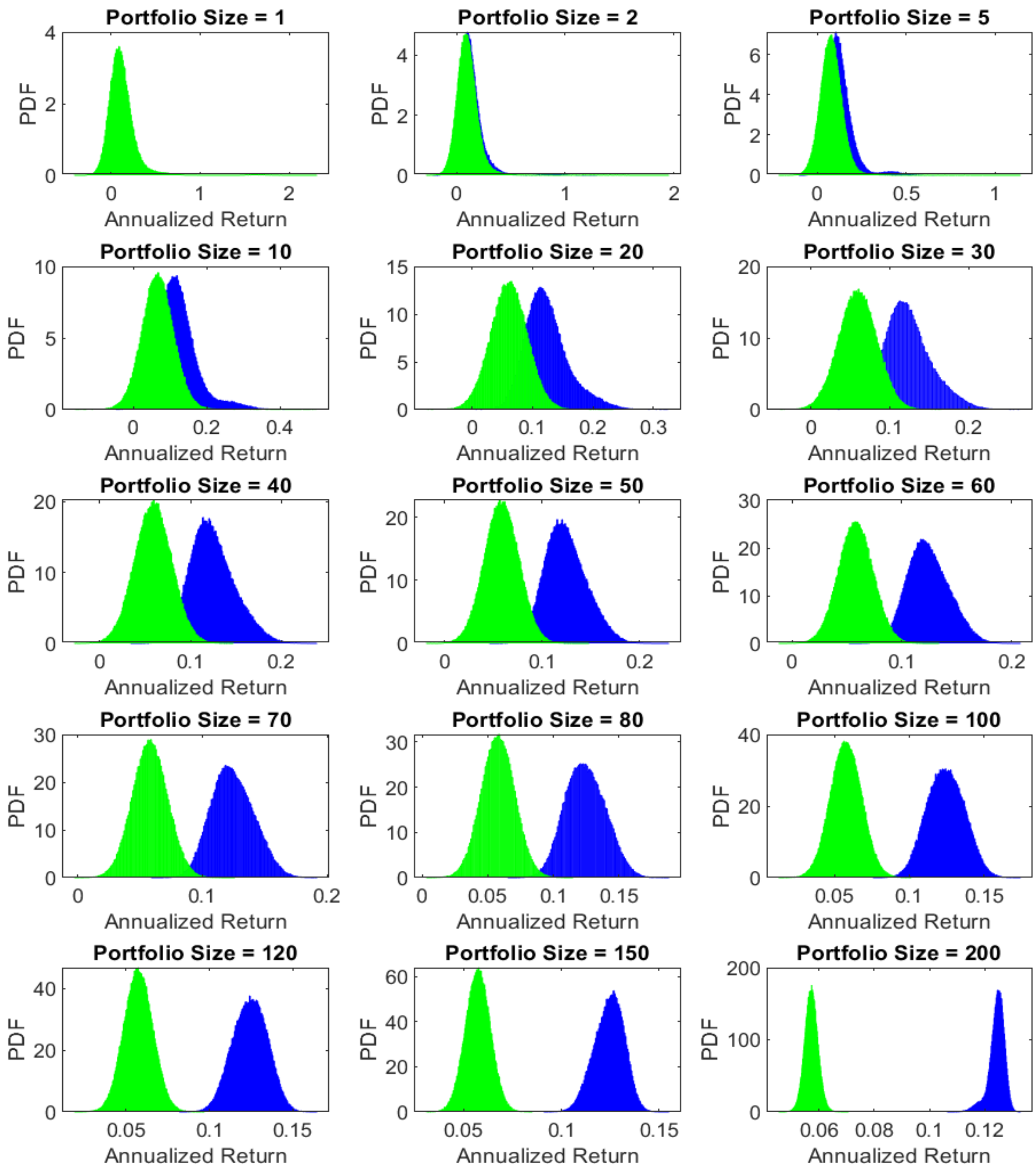

Figure 2: Return distributions for different portfolio sizes. This plot is based on the average of 10 years of yearly returns between 2011-01-03 and 2020-12-30. 200.000 simulations were performed. Blue distribution is for equally weighted portfolios and green distribution is for capitalization weighted portfolios.

Table 3: Percentiles of returns for equally weighted portfolios held for one year. 10 observations between 2011-01-03 and 2020-12-30 were averaged in this table. 200.000 simulations were run for each portfolio size.

Percentiles of	**Number of Stocks**															
Distribution	**1**	**2**	**5**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**100**	**120**	**150**	**200**	**212**
1%	-0.1383	-0.0677	-0.0039	0.0296	0.0559	0.0675	0.0755	0.0811	0.0856	0.0889	0.0919	0.0966	0.1007	0.1061	0.1151	0.1244
5%	-0.0688	-0.0192	0.0283	0.0534	0.0732	0.0823	0.0880	0.0922	0.0956	0.0981	0.1004	0.1041	0.1072	0.1113	0.1187	0.1244
10%	-0.0326	0.0076	0.0457	0.0667	0.0828	0.0902	0.0949	0.0984	0.1012	0.1031	0.1051	0.1083	0.1108	0.1142	0.1208	0.1244
25%	0.0301	0.0539	0.0770	0.0895	0.0994	0.1040	0.1071	0.1093	0.1110	0.1124	0.1135	0.1155	0.1171	0.1193	0.1230	0.1244
50%	0.1032	0.1088	0.1142	0.1173	0.1198	0.1211	0.1220	0.1227	0.1230	0.1234	0.1238	0.1243	0.1245	0.1249	0.1247	0.1244
75%	0.1858	0.1707	0.1567	0.1489	0.1437	0.1413	0.1396	0.1380	0.1368	0.1358	0.1348	0.1332	0.1317	0.1298	0.1263	0.1244
90%	0.2798	0.2425	0.2048	0.1867	0.1731	0.1640	0.1578	0.1531	0.1496	0.1470	0.1445	0.1407	0.1376	0.1338	0.1275	0.1244
95%	0.3606	0.3035	0.2458	0.2259	0.1943	0.1784	0.1687	0.1619	0.1570	0.1532	0.1499	0.1450	0.1410	0.1360	0.1283	0.1244
99%	0.6489	0.6092	0.4301	0.3032	0.2301	0.2026	0.1873	0.1773	0.1699	0.1642	0.1597	0.1525	0.1468	0.1402	0.1296	0.1244
Statistics																
Mean	0.1243	0.1244	0.1242	0.1243	0.1243	0.1243	0.1243	0.1243	0.1243	0.1243	0.1243	0.1244	0.1244	0.1244	0.1244	0.1244
Std.Dev.	0.1716	0.1209	0.0755	0.0529	0.0363	0.0289	0.0244	0.0211	0.0187	0.0168	0.0151	0.0125	0.0103	0.0076	0.0029	0.0000
Range	2.7165	1.5220	0.7479	0.4887	0.3193	0.2336	0.2030	0.1833	0.1550	0.1315	0.1222	0.0988	0.0820	0.0632	0.0265	0.0000
Shortfall %	1.5529	1.1542	0.7727	0.5708	0.4110	0.3386	0.2924	0.2588	0.2311	0.2114	0.1929	0.1627	0.1381	0.1055	0.0458	0

Table 4: Percentiles of returns for capitalization weighted portfolios held for one year. 10 observations between 2011-01-03 and 2020-12-30 were averaged in this table. 200.000 simulations were run for each portfolio size.

Percentiles of Distribution	**Number of Stocks**															
	1	**2**	**5**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**100**	**120**	**150**	**200**	**212**
1%	-0.1383	-0.0975	-0.0604	-0.0357	-0.0095	0.0037	0.0116	0.0174	0.0216	0.0255	0.0286	0.0334	0.0375	0.0426	0.0515	0.0572
5%	-0.0688	-0.0415	-0.0195	-0.0045	0.0115	0.0198	0.0255	0.0293	0.0323	0.0349	0.0370	0.0405	0.0433	0.0469	0.0533	0.0572
10%	-0.0326	-0.0118	0.0019	0.0112	0.0225	0.0286	0.0327	0.0358	0.0380	0.0399	0.0415	0.0443	0.0463	0.0492	0.0542	0.0572
25%	0.0301	0.0371	0.0368	0.0376	0.0409	0.0432	0.0450	0.0465	0.0475	0.0484	0.0492	0.0505	0.0516	0.0531	0.0556	0.0572
50%	0.1032	0.0935	0.0755	0.0660	0.0611	0.0594	0.0588	0.0584	0.0581	0.0579	0.0578	0.0576	0.0574	0.0574	0.0572	0.0572
75%	0.1858	0.1531	0.1151	0.0947	0.0813	0.0757	0.0725	0.0703	0.0687	0.0673	0.0663	0.0647	0.0633	0.0616	0.0588	0.0572
90%	0.2798	0.2155	0.1531	0.1212	0.0996	0.0904	0.0848	0.0811	0.0783	0.0760	0.0740	0.0710	0.0686	0.0655	0.0603	0.0572
95%	0.3606	0.2634	0.1787	0.1377	0.1107	0.0990	0.0923	0.0875	0.0840	0.0811	0.0786	0.0748	0.0718	0.0679	0.0612	0.0572
99%	0.6489	0.4206	0.2437	0.1709	0.1319	0.1155	0.1064	0.0996	0.0948	0.0905	0.0874	0.0821	0.0777	0.0722	0.0631	0.0572
Statistics																
Mean	0.1243	0.1036	0.0777	0.0663	0.0611	0.0594	0.0588	0.0584	0.0581	0.0579	0.0578	0.0576	0.0575	0.0574	0.0572	0.0572
Std.Dev.	0.1716	0.1202	0.0640	0.0436	0.0302	0.0241	0.0204	0.0177	0.0157	0.0140	0.0127	0.0105	0.0087	0.0064	0.0024	0.0000
Range	2.7165	2.2327	1.3580	0.6599	0.3084	0.2067	0.1743	0.1662	0.1453	0.1285	0.1112	0.0886	0.0779	0.0538	0.0252	0.0000
Shortfall %	2.2021	1.7256	1.3406	1.0788	0.7997	0.6532	0.5534	0.4874	0.4354	0.3893	0.3526	0.2917	0.2425	0.1805	0.0687	0

4.2 Five Year Periods

For the equally weighted portfolio, we performed Levene's test for testing the null hypothesis of equal variances across all portfolio sizes in table 5. The null is rejected due to a p-value of 0. Doing the ANOVA test for testing the null hypothesis of equal means across all portfolio sizes in table 5 yielded a p-value of 0.9986, so the null cannot be rejected. These results are as expected when looking at table 5.

Doing the same analysis for the capitalization weighted portfolio yielded a p-value of 0 for Levene's test and a p-value of 0 for the ANOVA test. Again, this is the expected results when looking at table 6.

From figure 3, it is seen that the equally weighted portfolios have two peaks (is bimodal). In section Appendix B.2 (p.61) it is shown that it is due to the small-and/or mid cap stocks.

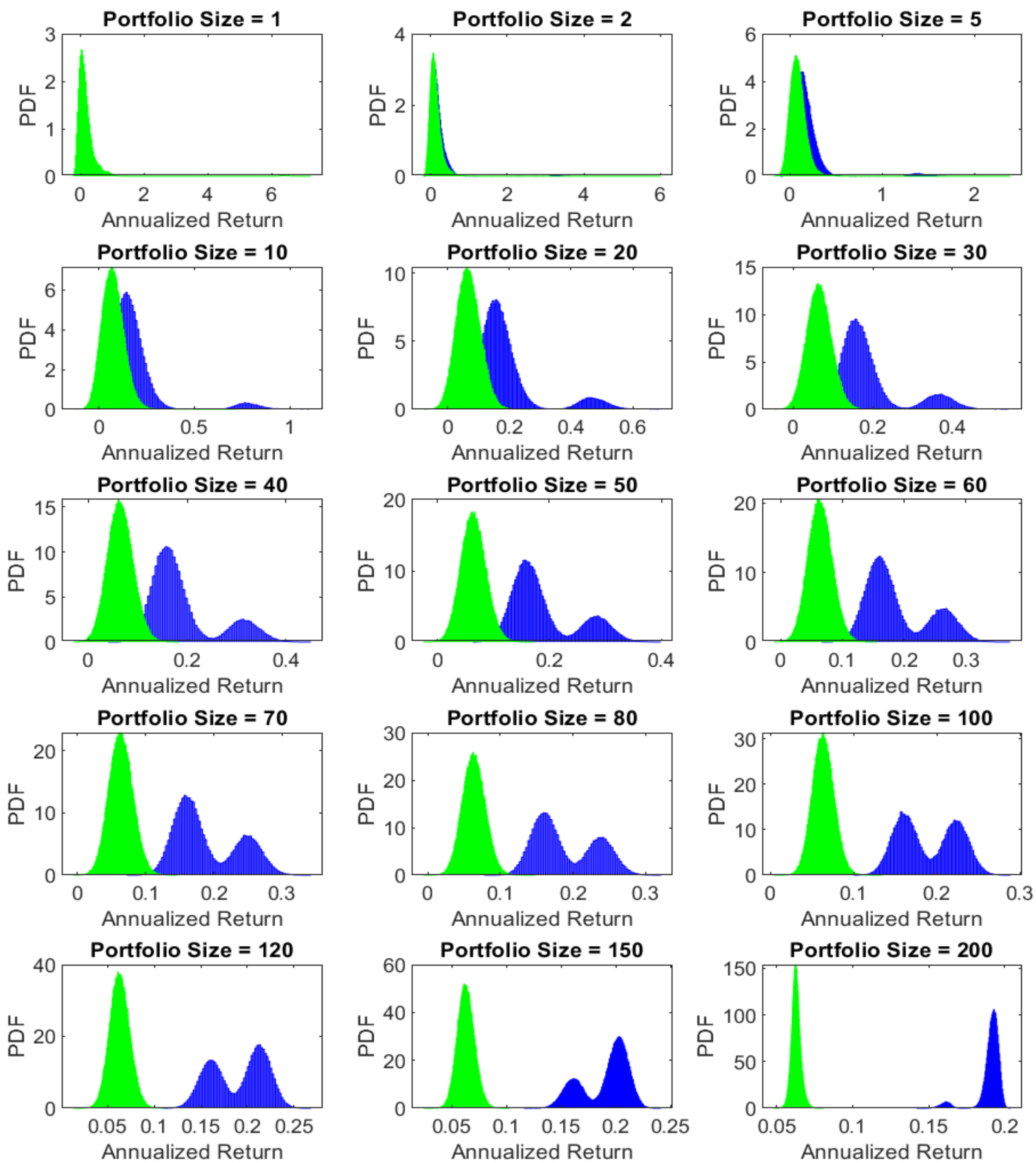

Figure 3: Return distributions for different portfolio sizes. This plot is based on the average of two five year returns between 2011-01-03 and 2020-12-30, returns have been annualized in the plot. 200.000 simulations were performed. Blue distribution is for equally weighted portfolios and green distribution is for capitalization weighted portfolios.

Table 5: Percentiles of returns for equal weighted portfolios held for five years, averaged over both periods. All numbers are annualized for ease of comparison. 200.000 simulations were run for each portfolio size.

Percentiles of	**Number of Stocks**															
Distribution	**1**	**2**	**5**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**100**	**120**	**150**	**200**	**212**
1%	-0.1372	-0.0796	-0.0112	0.0314	0.0664	0.0843	0.0964	0.1044	0.1107	0.1162	0.1212	0.1286	0.1350	0.1431	0.1574	0.1902
5%	-0.0864	-0.0322	0.0274	0.0626	0.0915	0.1056	0.1151	0.1213	0.1263	0.1307	0.1342	0.1403	0.1453	0.1517	0.1648	0.1902
10%	-0.0525	-0.0025	0.0506	0.0815	0.1062	0.1180	0.1257	0.1309	0.1352	0.1389	0.1418	0.1468	0.1513	0.1570	0.1854	0.1902
25%	0.0159	0.0540	0.0939	0.1156	0.1322	0.1399	0.1447	0.1483	0.1512	0.1537	0.1557	0.1595	0.1633	0.1705	0.1891	0.1902
50%	0.1143	0.1335	0.1518	0.1595	0.1644	0.1671	0.1690	0.1709	0.1725	0.1745	0.1766	0.1847	0.1976	0.1972	0.1919	0.1902
75%	0.2517	0.2413	0.2233	0.2115	0.2035	0.2029	0.2053	0.2184	0.2366	0.2354	0.2304	0.2219	0.2147	0.2060	0.1944	0.1902
90%	0.4452	0.3801	0.3050	0.2738	0.2847	0.3484	0.3132	0.2901	0.2730	0.2606	0.2504	0.2356	0.2249	0.2125	0.1963	0.1902
95%	0.6195	0.4798	0.3710	0.3728	0.4668	0.3816	0.3352	0.3064	0.2864	0.2713	0.2596	0.2424	0.2303	0.2161	0.1973	0.1902
99%	0.9861	0.8622	1.4152	0.8336	0.5294	0.4241	0.3673	0.3322	0.3077	0.2896	0.2756	0.2546	0.2396	0.2224	0.1991	0.1902
Statistics																
Mean	0.1904	0.1896	0.1900	0.1902	0.1902	0.1903	0.1901	0.1903	0.1899	0.1902	0.1899	0.1900	0.1903	0.1901	0.1902	0.1902
Std.Dev.	0.4792	0.3357	0.2115	0.1480	0.1018	0.0812	0.0681	0.0594	0.0523	0.0469	0.0422	0.0348	0.0288	0.0212	0.0081	0.0000
Range	7.4135	4.3039	1.9473	1.1572	0.6886	0.5094	0.4014	0.3334	0.3007	0.2666	0.2410	0.1871	0.1589	0.1237	0.0608	0.0000
Shortfall %	1.4545	1.1691	0.8561	0.6710	0.5187	0.4446	0.3947	0.3622	0.3358	0.3126	0.2946	0.2624	0.2360	0.2023	0.1337	0

Table 6: Percentiles of returns for capitalization weighted portfolios held for five years, averaged over both periods. All numbers are annualized for ease of comparison. 200.000 simulations were run for each portfolio size.

Percentiles of	**Number of Stocks**															
Distribution	**1**	**2**	**5**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**100**	**120**	**150**	**200**	**212**
1%	-0.1372	-0.1107	-0.0720	-0.0425	-0.0154	-0.0010	0.0085	0.0153	0.0204	0.0246	0.0284	0.0342	0.0391	0.0453	0.0560	0.0627
5%	-0.0864	-0.0614	-0.0330	-0.0129	0.0067	0.0170	0.0237	0.0287	0.0325	0.0355	0.0381	0.0423	0.0459	0.0503	0.0581	0.0627
10%	-0.0525	-0.0313	-0.0104	0.0041	0.0192	0.0271	0.0324	0.0362	0.0391	0.0414	0.0435	0.0468	0.0496	0.0530	0.0592	0.0627
25%	0.0159	0.0264	0.0305	0.0347	0.0408	0.0446	0.0472	0.0490	0.0505	0.0515	0.0526	0.0543	0.0558	0.0575	0.0608	0.0627
50%	0.1143	0.1018	0.0820	0.0718	0.0663	0.0648	0.0641	0.0638	0.0635	0.0632	0.0631	0.0629	0.0628	0.0627	0.0626	0.0627
75%	0.2517	0.1965	0.1384	0.1110	0.0933	0.0857	0.0816	0.0789	0.0769	0.0751	0.0739	0.0717	0.0701	0.0680	0.0645	0.0627
90%	0.4452	0.3152	0.1976	0.1484	0.1184	0.1053	0.0979	0.0928	0.0892	0.0860	0.0837	0.0798	0.0769	0.0730	0.0665	0.0627
95%	0.6195	0.4174	0.2395	0.1719	0.1332	0.1172	0.1077	0.1014	0.0967	0.0927	0.0896	0.0847	0.0810	0.0761	0.0679	0.0627
99%	0.9861	0.6923	0.3505	0.2211	0.1621	0.1405	0.1267	0.1176	0.1109	0.1052	0.1010	0.0943	0.0888	0.0820	0.0708	0.0627
Statistics																
Mean	0.1904	0.1374	0.0906	0.0747	0.0678	0.0657	0.0647	0.0643	0.0639	0.0635	0.0634	0.0631	0.0631	0.0629	0.0627	0.0627
Std.Dev.	0.4792	0.2416	0.0904	0.0569	0.0386	0.0305	0.0255	0.0221	0.0195	0.0174	0.0157	0.0129	0.0107	0.0078	0.0030	0.0000
Range	7.4135	6.1938	2.5532	0.8477	0.3455	0.2674	0.2111	0.1958	0.1661	0.1510	0.1522	0.1286	0.0966	0.0781	0.0324	0.0000
Shortfall %	2.3784	1.9795	1.5262	1.2060	0.8937	0.7296	0.6215	0.5429	0.4813	0.4332	0.3932	0.3251	0.2684	0.1978	0.0739	0

4.3 10 Year Periods

For the equally weighted portfolio, we performed Levene's test for testing the null hypothesis of equal variances across all portfolio sizes in table 7. The null is rejected due to a p-value of 0. Doing the ANOVA test for testing the null hypothesis of equal means across all portfolio sizes in table 7 yielded a p-value of 0.0371 so the null is rejected, which is reasonable when looking at table 7.

Doing the same analysis for the capitalization weighted portfolio, yielded a p-value of 0 for Levene's test and a p-value of 0 for the ANOVA test. Again, this is the expected results when looking at table 8.

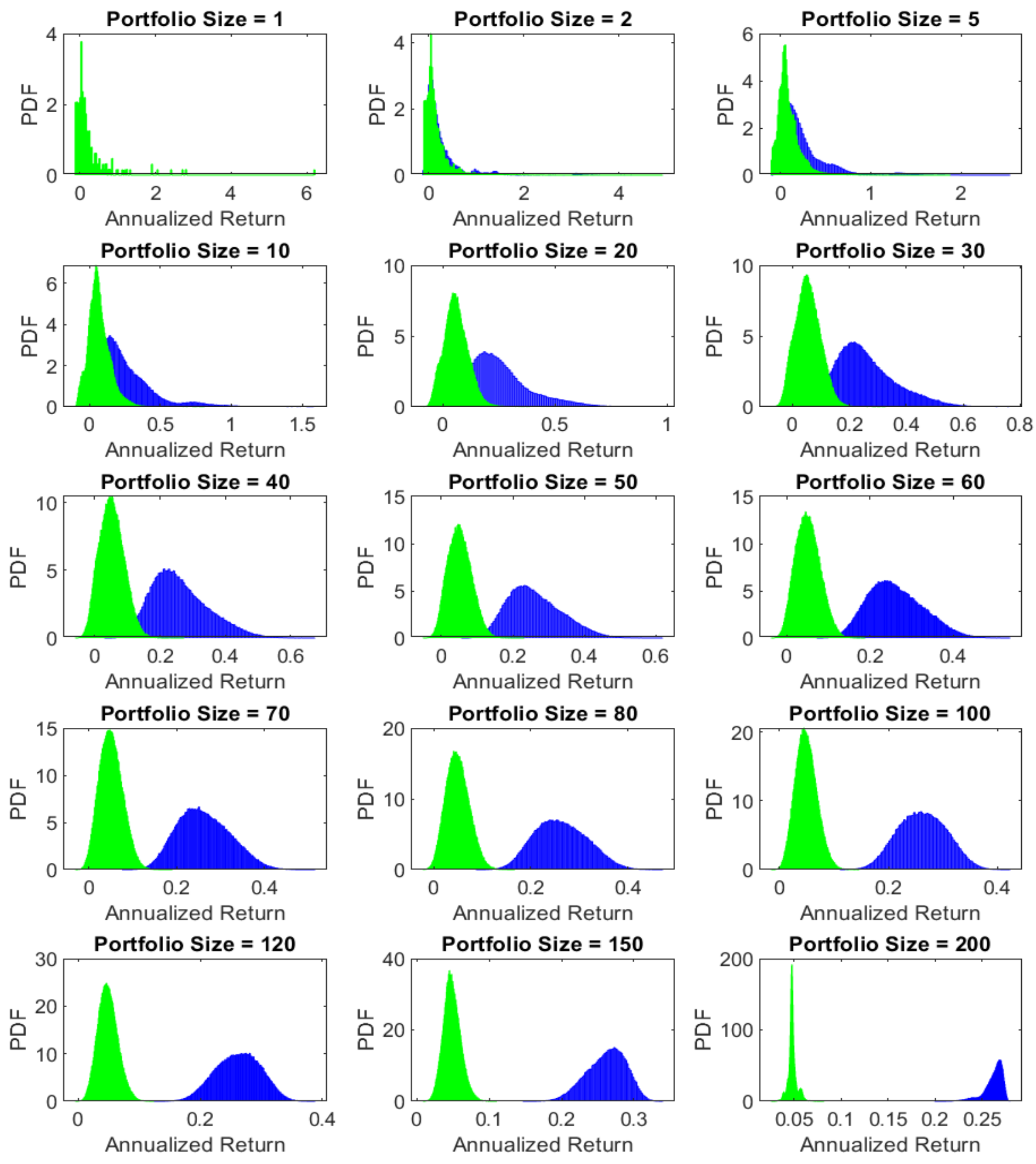

Figure 4: Return distributions for different portfolio sizes. This plot is based on one 10 year return between 2011-01-03 and 2020-12-30, returns have been annualized in the plot. 200.000 simulations were performed. Blue distribution is for equally weighted portfolios and green distribution is for capitalization weighted portfolios.

Table 7: Percentiles of returns for equally weighted portfolios held for ten years between 2011-01-03 and 2020-12-30. Annualized for ease of comparison. 200.000 simulations were run for each portfolio size.

Percentiles of	**Number of Stocks**															
Distribution	**1**	**2**	**5**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**100**	**120**	**150**	**200**	**212**
1%	-0.0993	-0.0809	-0.0196	0.0239	0.0675	0.0923	0.1099	0.1233	0.1349	0.1448	0.1534	0.1668	0.1798	0.1967	0.2269	0.2619
5%	-0.0926	-0.0406	0.0183	0.0607	0.1015	0.1252	0.1410	0.1535	0.1633	0.1717	0.1788	0.1907	0.2008	0.2146	0.2398	0.2619
10%	-0.0685	-0.0136	0.0426	0.0842	0.1234	0.1458	0.1606	0.1718	0.1805	0.1877	0.1940	0.2046	0.2133	0.2251	0.2489	0.2619
25%	0.0015	0.0417	0.0958	0.1338	0.1687	0.1868	0.1979	0.2062	0.2129	0.2181	0.2223	0.2299	0.2357	0.2437	0.2576	0.2619
50%	0.0976	0.1326	0.1837	0.2125	0.2358	0.2433	0.2486	0.2525	0.2553	0.2576	0.2590	0.2616	0.2629	0.2644	0.2644	0.2619
75%	0.2685	0.3104	0.3297	0.3376	0.3237	0.3199	0.3148	0.3107	0.3061	0.3032	0.2996	0.2938	0.2885	0.2817	0.2690	0.2619
90%	0.6761	0.6188	0.5785	0.4852	0.4453	0.4088	0.3838	0.3667	0.3537	0.3433	0.3342	0.3200	0.3088	0.2948	0.2719	0.2619
95%	1.1105	1.0459	0.7275	0.6732	0.5258	0.4621	0.4236	0.3989	0.3794	0.3654	0.3532	0.3343	0.3199	0.3018	0.2733	0.2619
99%	2.7099	1.9791	1.4038	0.9305	0.6539	0.5515	0.4914	0.4531	0.4251	0.4035	0.3857	0.3592	0.3386	0.3135	0.2755	0.2619
Statistics																
Mean	0.2598	0.2605	0.2622	0.2614	0.2618	0.2618	0.2616	0.2619	0.2618	0.2621	0.2618	0.2620	0.2619	0.2619	0.2619	0.2619
Std.Dev.	0.5974	0.4244	0.2681	0.1866	0.1289	0.1026	0.0862	0.0750	0.0661	0.0593	0.0534	0.0440	0.0364	0.0267	0.0102	0.0000
Range	6.2798	4.5879	2.6208	1.6391	0.9714	0.7332	0.6408	0.5607	0.4619	0.4355	0.3820	0.3090	0.2642	0.1925	0.0801	0.0000
Shortfall %	1.3537	1.1551	0.9301	0.7683	0.6122	0.5221	0.4615	0.4139	0.3766	0.3442	0.3171	0.2717	0.2332	0.1805	0.0844	0

Table 8: Percentiles of returns for capitalization weighted portfolios held for ten years between 2011-01-03 and 2020-12-30. Annualized for ease of comparison. 200.000 simulations were run for each portfolio size.

Percentiles of	**Number of Stocks**															
Distribution	**1**	**2**	**5**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**100**	**120**	**150**	**200**	**212**
1%	-0.0993	-0.0970	-0.0870	-0.0723	-0.0476	-0.0323	-0.0227	-0.0144	-0.0083	-0.0033	0.0012	0.0087	0.0148	0.0231	0.0368	0.0480
5%	-0.0926	-0.0762	-0.0562	-0.0432	-0.0261	-0.0143	-0.0058	0.0006	0.0052	0.0092	0.0126	0.0184	0.0232	0.0296	0.0407	0.0480
10%	-0.0685	-0.0548	-0.0267	-0.0170	-0.0101	-0.0024	0.0043	0.0096	0.0134	0.0169	0.0196	0.0244	0.0283	0.0335	0.0432	0.0480
25%	0.0015	0.0117	0.0167	0.0194	0.0217	0.0232	0.0252	0.0274	0.0292	0.0310	0.0325	0.0352	0.0372	0.0402	0.0460	0.0480
50%	0.0976	0.0843	0.0661	0.0596	0.0550	0.0523	0.0507	0.0498	0.0491	0.0487	0.0483	0.0481	0.0479	0.0476	0.0476	0.0480
75%	0.2685	0.2202	0.1461	0.1112	0.0921	0.0832	0.0774	0.0733	0.0702	0.0675	0.0653	0.0620	0.0592	0.0560	0.0495	0.0480
90%	0.6761	0.4505	0.2488	0.1675	0.1286	0.1122	0.1019	0.0950	0.0896	0.0851	0.0812	0.0752	0.0702	0.0642	0.0542	0.0480
95%	1.1105	0.6816	0.3242	0.2096	0.1502	0.1291	0.1164	0.1077	0.1011	0.0954	0.0906	0.0832	0.0771	0.0694	0.0573	0.0480
99%	2.7099	1.5207	0.5757	0.3044	0.1987	0.1622	0.1433	0.1314	0.1218	0.1140	0.1078	0.0980	0.0896	0.0796	0.0615	0.0480
Statistics																
Mean	0.2598	0.1706	0.0958	0.0701	0.0583	0.0543	0.0523	0.0513	0.0505	0.0500	0.0495	0.0491	0.0487	0.0484	0.0480	0.0480
Std.Dev.	0.5974	0.3012	0.1291	0.0775	0.0536	0.0437	0.0374	0.0327	0.0293	0.0262	0.0237	0.0196	0.0163	0.0120	0.0046	0.0000
Range	6.2798	5.0081	1.9752	0.9172	0.4733	0.4025	0.3323	0.2806	0.2246	0.2202	0.1883	0.1601	0.1283	0.1021	0.0557	0.0000
Shortfall %	2.9310	2.5879	2.1709	1.9010	1.5438	1.2988	1.1206	0.9875	0.8925	0.8075	0.7372	0.6155	0.5173	0.3820	0.1508	0

4.4 Comparison Between Periods

In this section, differences between different holding periods is studied. Due to the Equally weighted portfolios being better in general, I focus the analysis on that and only provide a plot (figure 6) of the capitalization weighted portfolios for the interesterd reader.

From figure 5 (and a CDF plot not presented here), it is clear that for portfolio sizes of 5 and greater we have that:

- Portfolios held for 10 years second-order stochastically dominates portfolios held for five years and one year respectively.
- Portfolios held for five years second-order stochastically dominates portfolios held for one year.

This implies that an risk averse, utility maximizing investor, would ceteris paribus prefer the portfolios held for 10 years.

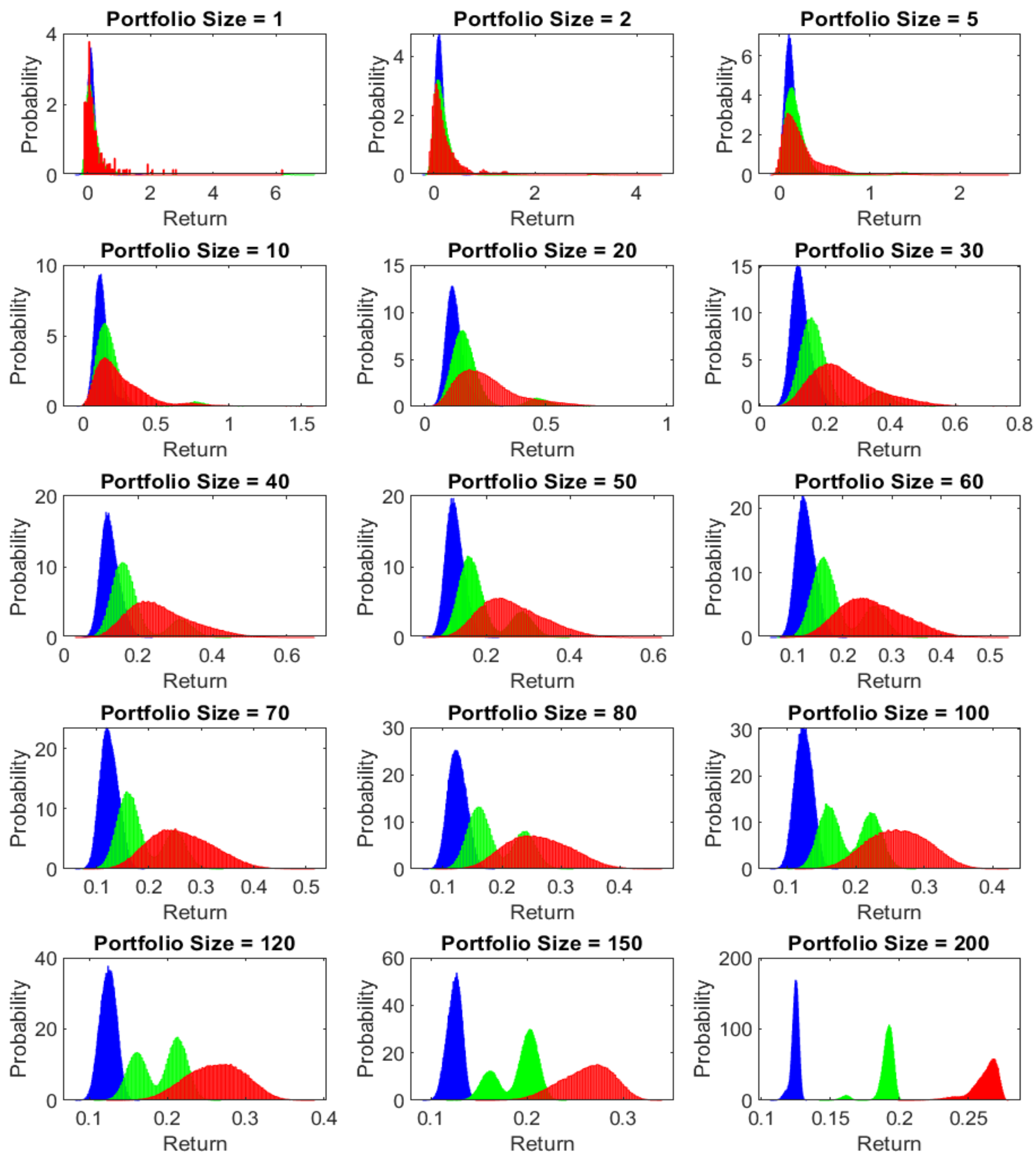

Figure 5: Return distributions for equally weighted portfolios based on 200.000 simulations between 2011-01-03 and 2020-12-30. Blue plot is averaged across ten yearly returns. Green plot is averaged across two five year returns and red plot is for the entire 10 year period containing only one observation. All returns are annualized for ease of comparison.

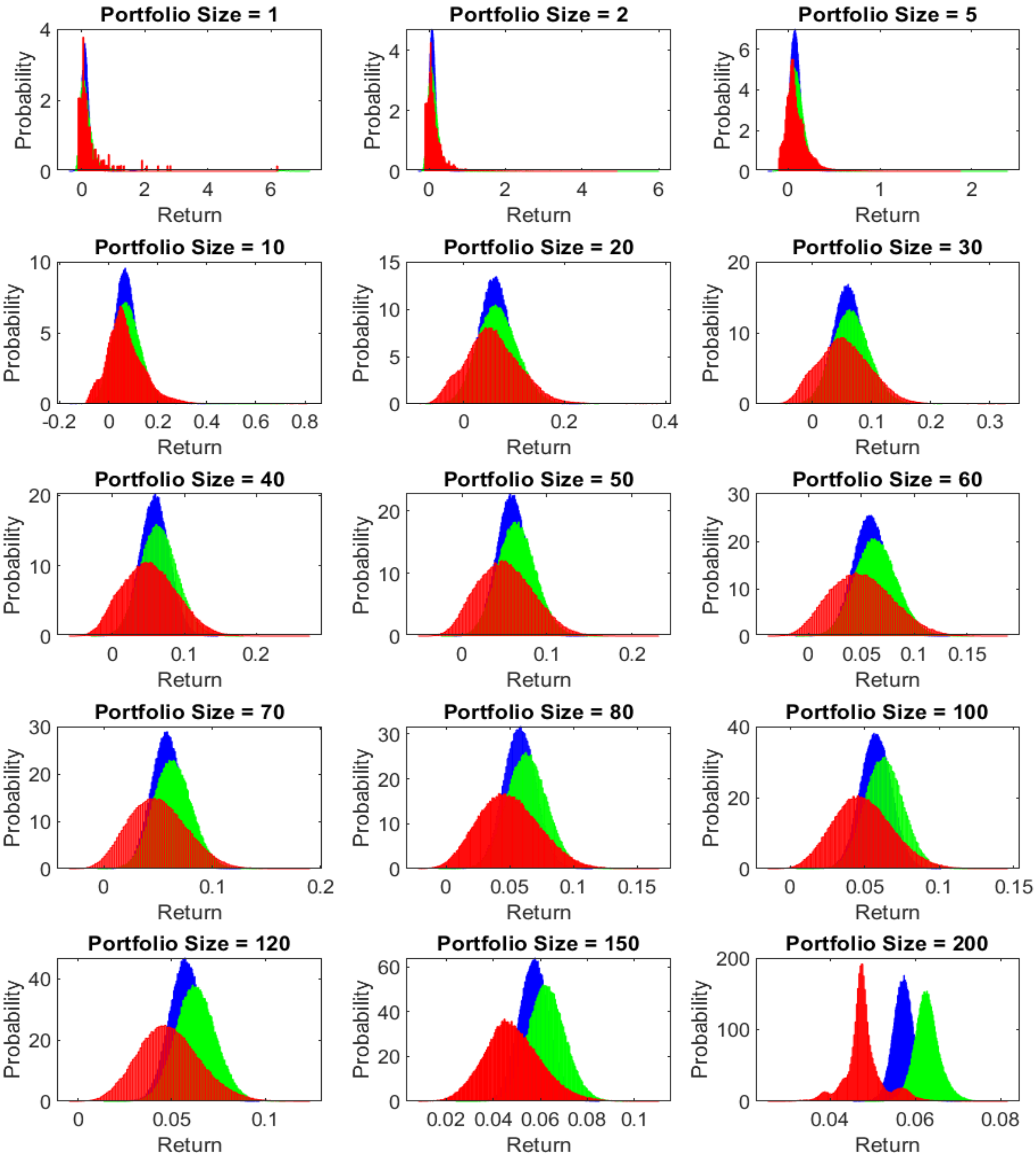

Figure 6: Return distributions for capitalization weighted portfolios based on 200.000 simulations between 2011-01-03 and 2020-12-30. Blue plot is averaged across ten yearly returns. Green plot is averaged across two five year returns and red plot is for the entire 10 year period containing only one observation. All returns are annualized for ease of comparison.

Chapter 5

Conclusions and Further Research

Professional investors and Finance academics are in general agreement that investors should hold diversified portfolios to some extent. However the specific question of *"How many stocks does it take to be well-diversified"* is less clear cut and is obviously strongly influenced by the preferences of the individual investor.

If one major concern with not holding diversified portfolios is the possibility of returns ending below the stock universe studied, the shortfall risk measure is natural to use. For equally weighted portfolios, reducing the shortfall risk to 20% for a 10 year investment period would take around 120-150 stocks. For five (one) year periods, it would take 150 stocks (70-80 stocks). For private investors, owning that many stocks is most likely not reasonable[1], especially if smaller amounts of money is invested. The solution to beeing well-diversified would then simply be to own some investment funds in the portfolio. If the investor wants some specific stocks they can easily be added to the portfolio.

It was seen by visual inspection, that in general, equally weighted portfolios second-order stochastically dominated capitalization weighted portfolios for portfolio sizes of 5 and greater. The conclusion is however not completly

[1] One obvious reason is transaction costs such as brokerage fees, searching costs (which stocks to buy?) and monitoring costs (what is happening with my stocks?).

waterproof due to the market capitalizations used for different dates was an approximation. See chapter 3 for details. Similarly, portfolios held for 10 years second-order stochastically dominates portfolios held for five years and one year respectively for portfolio sizes of five and greater. This conclusions however should be interpreted with care as there is a literature and research available on the topic of "Time Diversification", it deserves to be evaluated more thoroughly before a general conclusion or guideline is taken. The conclusions however holds for the data set studied in this thesis and at least provides an indication of what we could expect. The concept of stochastic dominance could be used in further studies to see what portfolio sizes dominates other. In connection to this, if the results are robust through time with changing market conditions would be interesting to see. It is also possible to do formal hypothesis tests for stochastic dominance (contrary to a visual inspection as was done in this thesis), see Whang [2019] for details. This could be done in further research.

Chapter 4 contains illustrative tables. For instance, the tables quantifies the simple fact that owning more stocks make the worst case scenarios (for instance measured by the first or fifth percentile) better. Similarly, the other side of the coin is that in general (exception for table 5 whose distribution plot had two peaks, see figure 3 on p.35) the best case scenario gets worse with more stocks. So one could say that the price to pay for decreasing the downside risk is by decreasing the upside potential as well. This is easy to establish in intuitive terms but the tables quantifies it which provides further insight into the concept of portfolio diversification.

Is it possible to reconcile the seemingly paradoxical fact that in Sweden year 2020, the average number of stocks held per shareholder is 4.5 and 41.79% held only one stock [2]? The average market value of the portfolios were around SEK 523.000 [Euroclear, 2020, p.12, p.14]. As a proxy [3], in 2016 the average market value of portfolios were SEK 517.000 and the median portfolio was worth SEK 30.000. The big difference between the mean- and median portfolio value is explained by skewed ownership where a small number of shareholders own portfolios worth much more than the average. For

[2] Many people are in practice more diversified by for instance having invested in real estate or through part of their pension savings which are in some funds.

[3] I use this proxy since data from 2020 regarding the information to come, to my knowledge, is not available.

instance, the five percent with the largest holdings owned around 79 percent of the share wealth SCB [2016]. With these, relatively smaller amounts [4] it is less surprising that 41.79% of the Swedish shareholders own only one stock since even if they lose the entire investment, it will probably not put them in a catastrophic financial situation whereas it could potentially lead to major gains as has been quantified in the tables from chapter 4. For a discussion of these matters in a behavioral finance framework, see for instance Statman [2004]. In brief, he theorize the fact that people are willing to take high risk with only part of their money due to the desire for riches. For some this could mean that they gamble at a casino and for others that they buy one stock that they believe could be the next "winner". As long as people do know what risk they are undertaking, this is a question of individual preferences. Do people know what risks they are taking?

Further studies would benefit by handling the biases in this thesis such as the survivalship bias, approximation of market capitalizations and including dividends.

Diversification is in practice a multi-dimensional problem and to diversify in the "usual meaning" of the word can be done in more ways than just buying more stocks [5]. To name a few examples it is possible to: diversify by holding different asset classes (such as bonds or gold), diversify through time and many different weighting schemes affecting the final outcome is possible to use. These other dimensions of "Diversification" would be fruitful to study.

No doubt, much work has been done in the field of portfolio diversification but I believe there is more to do in order to understand the multi-dimensional complexity of it.

[4] To put SEK 30.000 in a context, the average income for people in sweden were SEK 35.300 year 2019 SCB [2021].

[5] Imagine an extreme crisis such as a world war. Then perhaps, investing in canned food and water would be better then holding 100 different stocks? Should these kind of thoughts be incorporated in a formal analysis and if yes, how could it be done in a modelling framework? If not included, "Diversification" would perhaps not help you when you need it at most.

Appendix A

Derivation of Optimal Weights

The contents in this section together with related notions, can be found in chapter 3 of the book by Petters and Dong [2016] which is very readable and at a mathematical level.

In this section we will derive the solution to the optimization problem given by equation (2.1) subject to the constraints given by equation (2.2) and (2.3). First we present some preliminaries that will be needed.

Lemma 16. *The inverse of a invertible, symmetric matrix is also symmetric, i.e.* $\mathbf{A}^{-1} = (\mathbf{A}^{-1})^T$.

Proof. Proof is omitted. □

Assumption 17. *The covariance matrix of return rates* R, $\mathbf{\Sigma}$ *is invertible.* □

Proposition 18. *Assumption 17 implies that there is no redundant security in the portfolio in the sense that no security return is a linear combination of the others.*

Proof. Assume on the contrary, that the first security return is a linear combination of the other security returns:

$$R_1 = a_2 R_2 + ... + a_N R_N$$

The covariance matrix $\mathbf{\Sigma}$ can be written as:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1N} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \dots & \sigma_{NN} \end{bmatrix}$$

where as usual $\sigma_{ij} = Cov(R_i, R_j) = \sigma_{ji}$. Expanding the first entry in the first column of the covariance matrix $\mathbf{\Sigma}$, we get:

$$\begin{aligned} \sigma_{11} = Cov(R_1, R_1) =& Cov(R_1, a_2R_2 + ... + a_NR_N) \\ =& a_2Cov(R_1, R_2) + ... + a_NCov(R_1, R_N) \\ =& a_2\sigma_{12} + \dots + a_N\sigma_{1N} \end{aligned}$$

Analogous calculations are done for the other entries in the first column yielding the following equality:

$$\begin{bmatrix} \sigma_{11} \\ \sigma_{21} \\ \vdots \\ \sigma_{N1} \end{bmatrix} = \begin{bmatrix} a_2\sigma_{12} + \dots + a_N\sigma_{1N} \\ a_2\sigma_{22} + \dots + a_N\sigma_{2N} \\ \vdots \\ a_2\sigma_{N2} + \dots + a_N\sigma_{NN} \end{bmatrix}$$

Denoting column i in $\mathbf{\Sigma}$ by $\mathbf{c}_i$, we see that the first column $\mathbf{c}_1$ is a linear combination of the others:

$$\mathbf{c}_1 = a_2\mathbf{c}_2 + \dots + a_N\mathbf{c}_N$$

This for instance imply that the determinant of Σ is 0 and hence non-invertible.

Arguing by contradiction, we have shown that the assumption of an invertible covariance matrix, implies that no secuirty return is a linear combination of the others. □

Proposition 19. *The covariance matrix, $\mathbf{\Sigma}$, of security return rates R is positive definite.*

Proof. A covariance matrix is always positive semidefinite. This follows from the fact that:

$$\mathbf{x}^T \mathbf{\Sigma} \mathbf{x} = Var(x_1 R_1 + \ldots + x_N R_N) \geq 0$$

We assumed previously that the covariance matrix, $\mathbf{\Sigma}$, is invertible. This implies that the determinant, $det(\mathbf{\Sigma}) > 0$. Since the determinant is equal to the product of its eigenvalues, $det(\mathbf{\Sigma}) > 0$ implies that all eigenvalues are greater than 0. Finally, in Linear Algebra there is a theorem stating that a matrix is positive definite iff all eigenvalues are positive. Invoking this theorem shows that the covariance matrix of security return rates is positive definite. This concludes the proof. □

Lemma 20.

$$\frac{\partial \mathbf{x^T a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{a^T x}}{\partial \mathbf{x}} = \mathbf{a}. \tag{A.1}$$

Proof. Proof is omitted. □

Remark 21. *The formula presented in this lemma is simple to remember by using the mnemonic technique of comparing it to the single variable case. For example, if a is a scalar and $f(x) = ax$, then $f'(x) = a$.* □

We next present computational rules for the gradient and hessian of a scalar valued function of several variables when the function is of a specific form.

Lemma 22 (The Gradient)**.** *If $f(\mathbf{x}) = \mathbf{x^T A x}$ is a real valued function of several variables, i.e. $f : \mathbb{R}^n \to \mathbb{R}$, where $\mathbf{A}$ is a $n \times n$ real matrix. Then the gradient of f, $\nabla f = \frac{\partial f}{\partial \mathbf{x}}$, is defined as:*

$$\nabla f = \frac{\partial f}{\partial \mathbf{x}} = \frac{\partial (\mathbf{x^T A x})}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A^T})\mathbf{x}. \tag{A.2}$$

Proof. Proof is omitted. □

Lemma 23 (The Hessian)**.** *If* $f(\mathbf{x}) = \mathbf{x}^{\mathbf{T}}\mathbf{A}\mathbf{x}$ *is a real valued function of several variables, i.e.* $f : \mathbb{R}^n \to \mathbb{R}$*. Then the hessian of* f, $\frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^{\mathbf{T}}}$, *is defined as:*

$$\frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^{\mathbf{T}}} = \frac{\partial^2 (\mathbf{x}^{\mathbf{T}}\mathbf{A}\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^{\mathbf{T}}} = \mathbf{A} + \mathbf{A}^{\mathbf{T}}. \tag{A.3}$$

Proof. Proof is omitted. □

We now have the tools needed to solve the portfolio minimization problem. The method of Lagrangian multipliers will be used and it is assumed the reader has a working knowledge of it.

Problem 24. *We solve the portfolio minimization problem by finding the portfolio weight vector* $\mathbf{w}$ *that solves:*

$$\mathbf{w} = \arg\min_{\mathbf{w}} \frac{\mathbf{w}^{\mathbf{T}}\boldsymbol{\Sigma}\mathbf{w}}{2} \tag{A.4}$$

subject to the constraints:

$$\mathbf{w}^{\mathbf{T}}\boldsymbol{\mu} = \mu^* \tag{A.5}$$

$$\mathbf{w}^{\mathbf{T}}\mathbf{1} = \sum_i w_i = 1 \tag{A.6}$$

Remark 25. *Notice that the objective function as given by equation (A.4) is divided by 2 compared to the original problem in equation (2.1) at (p.3). This does not matter since the solutions are equivalent. Also, the standard deviation could be used as risk measure because in an optimization context it is equivalent to the variance. But the variance is much simpler to work with in this case.*

Solution. We begin by setting up the Lagrangian, $\mathcal{L}$:

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}) = \frac{\mathbf{w^T \Sigma w}}{2} + \lambda_1(1 - \mathbf{w}^T\mathbf{1}) + \lambda_2(\mu^* - \mathbf{w}^T\boldsymbol{\mu}) = f(\mathbf{w}) + \boldsymbol{\lambda^T}\mathbf{h}(\mathbf{w}) \tag{A.7}$$

where we defined,

$$f(\mathbf{w}) = \frac{\mathbf{w^T \Sigma w}}{2}$$

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}$$

$$\mathbf{h}(\mathbf{w}) = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} 1 - \mathbf{w}^T\mathbf{1} \\ \mu^* - \mathbf{w}^T\boldsymbol{\mu} \end{bmatrix}$$

Using the lemma 20 and lemma 22, we get the following two first order conditions:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}}(\mathbf{w}, \boldsymbol{\lambda}) = \frac{(\mathbf{\Sigma} + \mathbf{\Sigma^T})}{2}\mathbf{w} - \lambda_1\mathbf{1} - \lambda_2\boldsymbol{\mu} = \mathbf{\Sigma w} - \lambda_1\mathbf{1} - \lambda_2\boldsymbol{\mu} \tag{A.8}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} = \mathbf{h}(\mathbf{w}) \tag{A.9}$$

In the last equality of equation (A.8) we used that the covariance matrix $\mathbf{\Sigma}$ is symmetric, i.e. $\Sigma^T = \Sigma$ which implies that $(\mathbf{\Sigma} + \mathbf{\Sigma}^T)/2 = \mathbf{\Sigma}$.

To find critical points, the first order conditions are set equal to 0. We first obtain from equation (A.8):

$$\begin{aligned} \mathbf{\Sigma w} = \lambda_1\mathbf{1} + \lambda_2\boldsymbol{\mu} &\iff \mathbf{w} = \lambda_1\mathbf{\Sigma}^{-1}\mathbf{1} + \lambda_2\mathbf{\Sigma}^{-1}\boldsymbol{\mu} \\ &\iff \mathbf{w}^T = \lambda_1\mathbf{1}^T\mathbf{\Sigma}^{-1} + \lambda_2\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1} \end{aligned} \tag{A.10}$$

In the first equivalence we used assumption 17 of $\mathbf{\Sigma}$ beeing invertible. In the last equivalence we used the property of lemma 16 that the inverse of a symmetric matrix is still symmetric, i.e. $(\boldsymbol{\Sigma}^{-1})^T = \boldsymbol{\Sigma}^{-1}$. Setting the components of $\mathbf{h}(\mathbf{w}) = 0$, then we secondly get:

$$1 = \mathbf{w}^T\mathbf{1} \tag{A.11}$$
$$\mu^* = \mathbf{w}^T\boldsymbol{\mu} \tag{A.12}$$

Plugging in the expression for $\mathbf{w}^T$ from equation (A.10) into the two equations above, we obtain:

$$1 = \mathbf{w}^T\mathbf{1} = \lambda_1\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1} + \lambda_2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{1} \tag{A.13}$$
$$\mu^* = \mathbf{w}^T\boldsymbol{\mu} = \lambda_1\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \lambda_2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{A.14}$$

Define $A = \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}$, $B = \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}$ and $C = \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.

Notice that B is a scalar (check the dimensions) and therefore $B^T = B$. Equipped with this notation, we get that:

$$\begin{bmatrix} A & B \\ B & C \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 1 \\ \mu^* \end{bmatrix}$$

Denote

$$\begin{bmatrix} A & B \\ B & C \end{bmatrix} := K$$

Then,

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = K^{-1} \begin{bmatrix} 1 \\ \mu^* \end{bmatrix} = \frac{1}{AC - B^2} \begin{bmatrix} C & -B \\ -B & A \end{bmatrix} \begin{bmatrix} 1 \\ \mu^* \end{bmatrix}$$

Identifying components, we get:

$\lambda_1 = \frac{C - \mu^* B}{AC - B^2}$ and $\lambda_2 = \frac{\mu^* A - B}{AC - B^2}$.

Plugging in the multipliers (λ_1 and λ_2) in equation (A.10), we then get:

$$\mathbf{w}^T = \frac{C - \mu^* B}{AC - B^2}\mathbf{1}^T\mathbf{\Sigma}^{-1} + \frac{\mu^* A - B}{AC - B^2}\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1}$$

Taking the transpose to obtain the solution $\mathbf{w}$, we finally get:

$$\mathbf{w} = \frac{C - \mu^* B}{AC - B^2}\mathbf{\Sigma}^{-1}\mathbf{1} + \frac{\mu^* A - B}{AC - B^2}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$$

It remains to show that this candidate solution indeed is a minimum which will be done by invoking the second derivative test. Using Lemma 23 for calculating the hessian, we get that:

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w} \partial \mathbf{w^T}}(\mathbf{w}, \boldsymbol{\lambda}) = \mathbf{\Sigma}$$

It was shown in proposition 19 that the hessian (i.e. the covariance matrix $\mathbf{\Sigma}$ in this case) is positive definite. Hence it is a minimum point.

□

Appendix B

Results and Analysis - More Details

In chapter 4 the plots and tables are constructed by averaging individual tables. In the case of one year returns, ten observations between [2011-01-03] and [2020-12-30] were averaged. In the case of five year returns, two observations were averaged and in the case of 10 year returns there was only one observation. This gives an overview and simplifies analysis of more general character. Nevertheless, the reader might get the false impression that *all* periods were similar to these aggregated results. This is not the case and I will provide some comments that are interesting for the individual observations, not averaged. For instance, the distributions are changing both in location and shape.

The takeaway for the reader should be that we cannot expect too much "regularity" from one year to another making it a dangerous endeavour to generalize the results without care.

B.1 One Year Periods

In interest of brevity not all plots are presented. Some interesting observations from these plots were:

- In 2011 (see figure 7) you most likely would get a negative return. For larger portfolio sizes you surely would. This kind of results can get the individual investor to think it is "safer" to invest in few stocks. Only one year later (see figure 8), owning few stocks can lead to substantial negative returns while owning many stocks surely lead to positive returns. What one prefers is a matter of preference but it is important to know that owning few stocks, in general, is riskier in terms of the downside risk.

- In 2012 (see figure 8) for portfolio sizes greater than 30, the cap weighted portfolios second-order stochastically dominates the equally weighted portfolios as seen by a visual inspection.
- In 2015 (see figure 9) the equally weighted portfolio is bimodal for larger portfolio sizes.
- In 2018 (see figure 10) we again see the bimodal characteristic of the equally weighted portfolio and there is a high risk of getting a negative return.

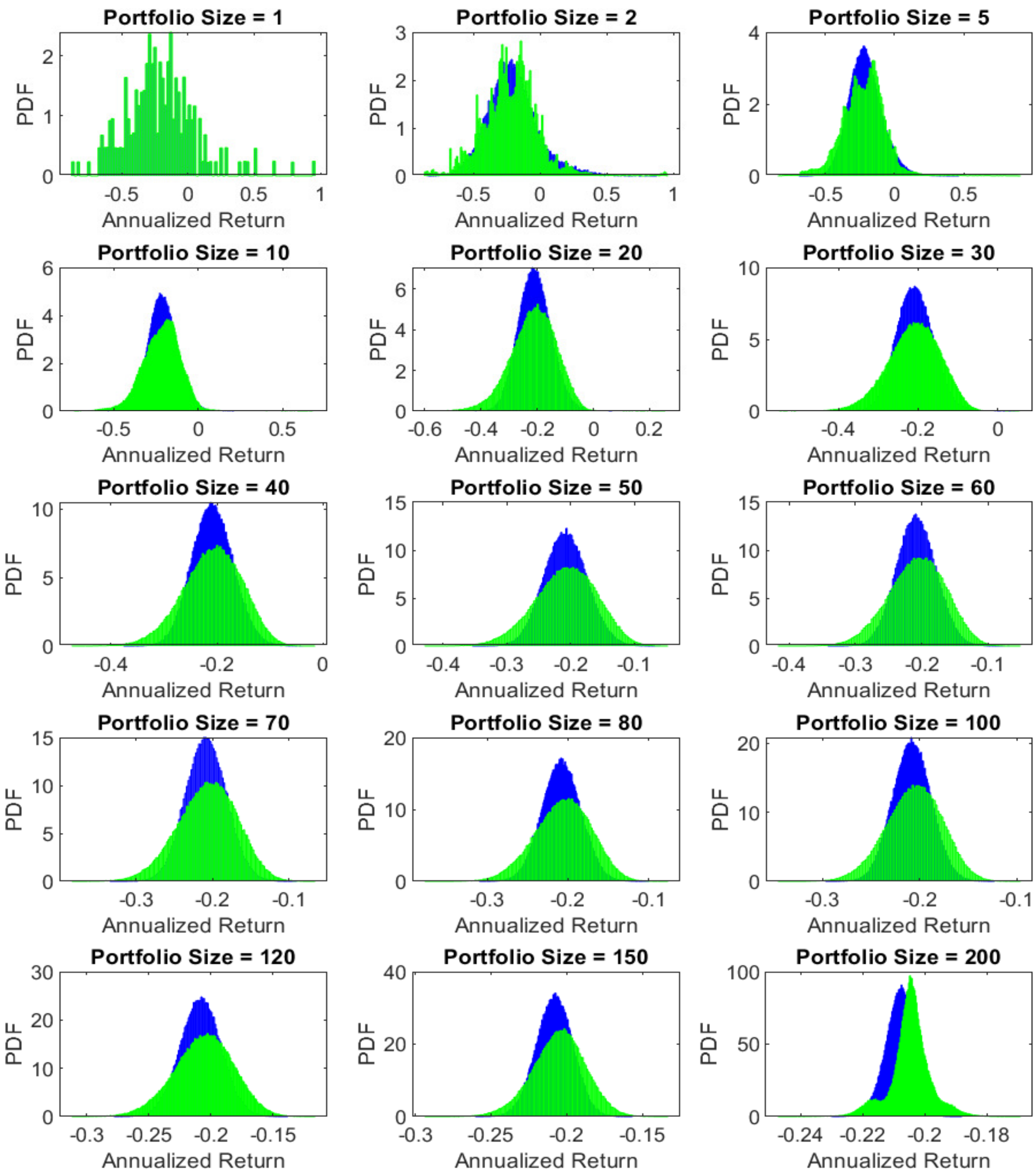

Figure 7: Return distributions for different portfolio sizes. This plot is based on 200.000 simulations between 2011-01-03 and 2011-12-30. Blue distribution is for equally weighted portfolios and green distribution is for capitalization weighted portfolios. Notice how positive returns only are obtainable for smaller portfolio sizes.

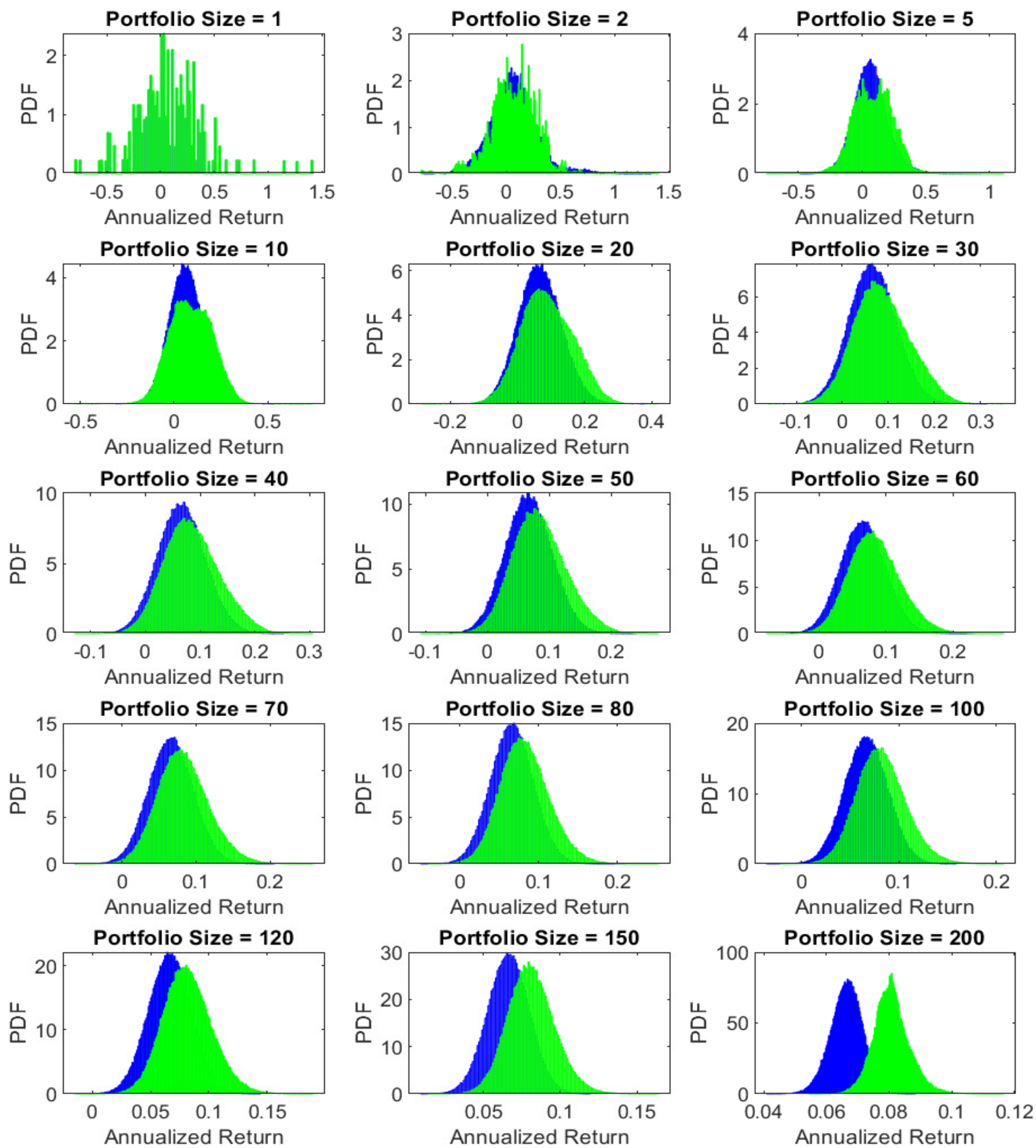

Figure 8: Return distributions for different portfolio sizes. This plot is based on 200.000 simulations between 2012-01-02 and 2012-12-28. Blue distribution is for equally weighted portfolios and green distribution is for capitalization weighted portfolios. Notice how the capitalization weighted distribution in general is better for portfolio sizes of 30 or greater.

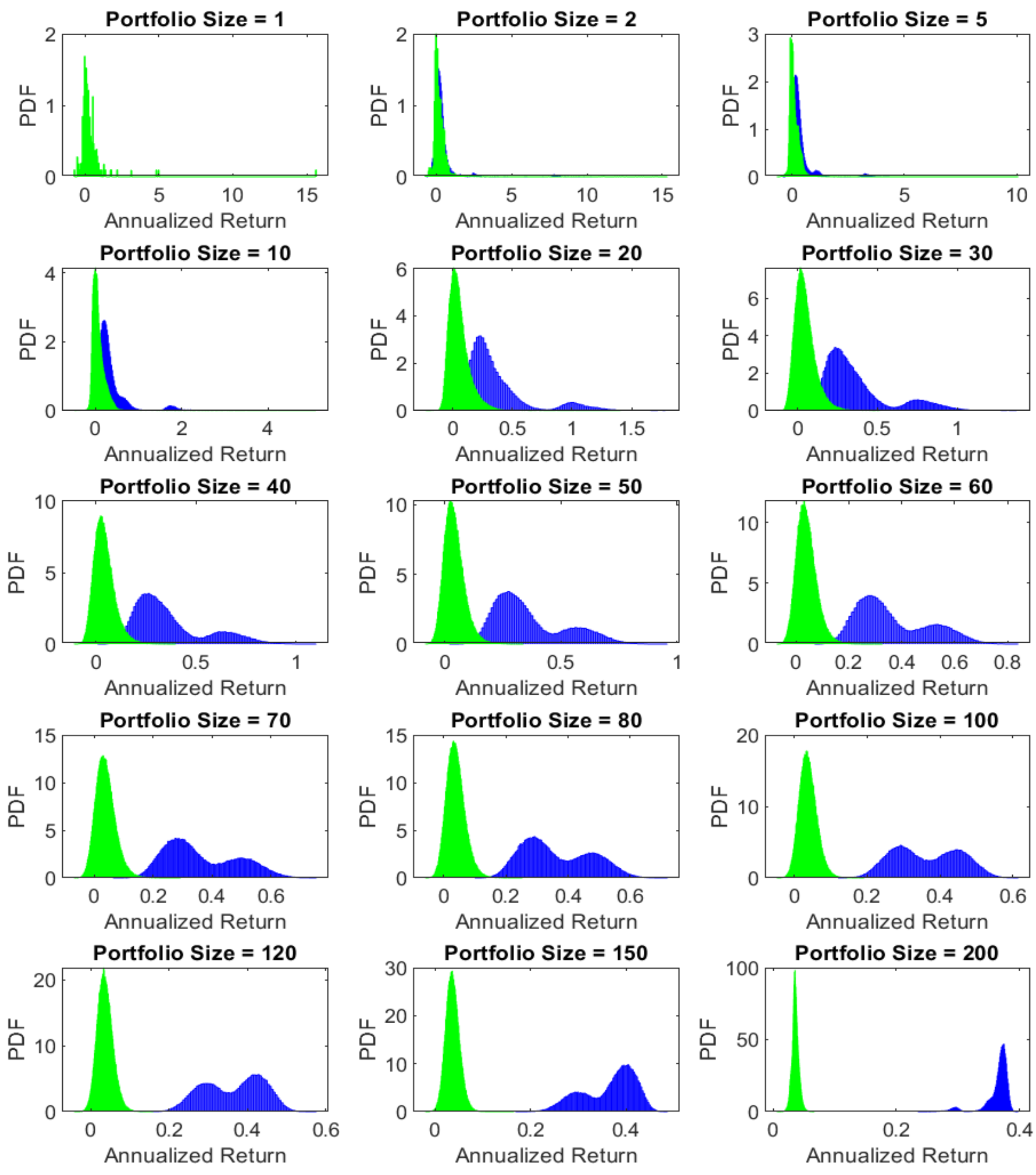

Figure 9: Return distributions for different portfolio sizes. This plot is based on 200.000 simulations between 2015-01-02 and 2015-12-30. Blue distribution is for equally weighted portfolios and green distribution is for capitalization weighted portfolios. Notice how the equally weighted distribution is bimodal.

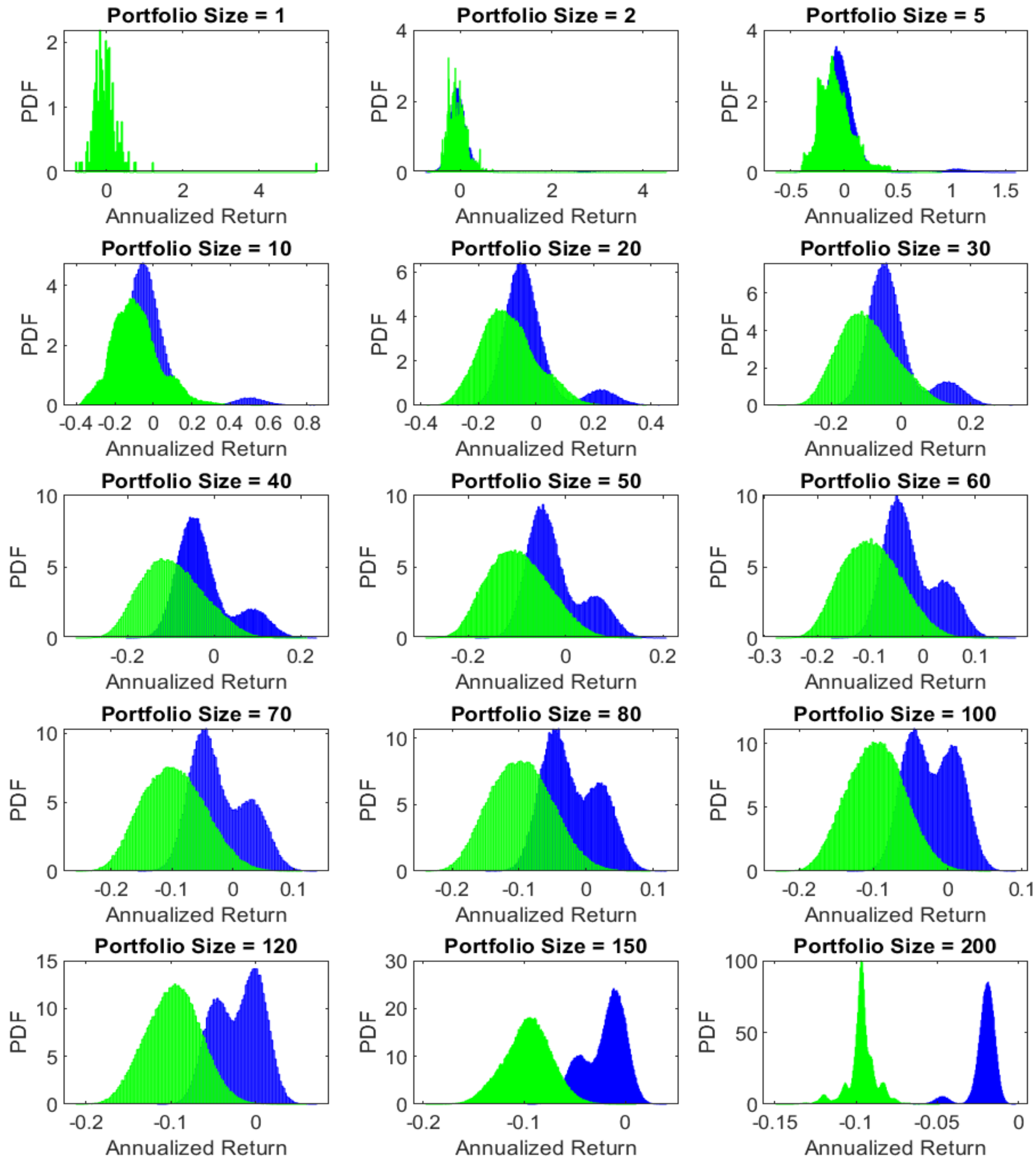

Figure 10: Return distributions for different portfolio sizes. This plot is based on 200.000 simulations between 2018-01-02 and 2018-12-28. Blue distribution is for equally weighted portfolios and green distribution is for capitalization weighted portfolios. Notice how the equally weighted distributions are bimodal and negative returns are likely, especially for smaller portfolio sizes.

B.2 Five Year Periods

Figure 3 is the average of two five year periods. Looking at the periods seperately, we see that the distributions have changed considerably. Figure 11 is bimodal for larger portfolio sizes. To understand this characteristic better, the same plot was produced but only containing large cap stocks. The result, presented in Figure 13 shows that there is no bimodal property for this market. Hence, we can infer that the bimodal property is due to inclusion of small- and/or mid cap stocks who had higher returns than the large cap stocks only. This data set, contained 99 stocks (compare with 212 before) and hence different portfolio sizes were looked at. For example, the portfolio size of 200 were not possible to study in the data set consisting of large cap stocks only.

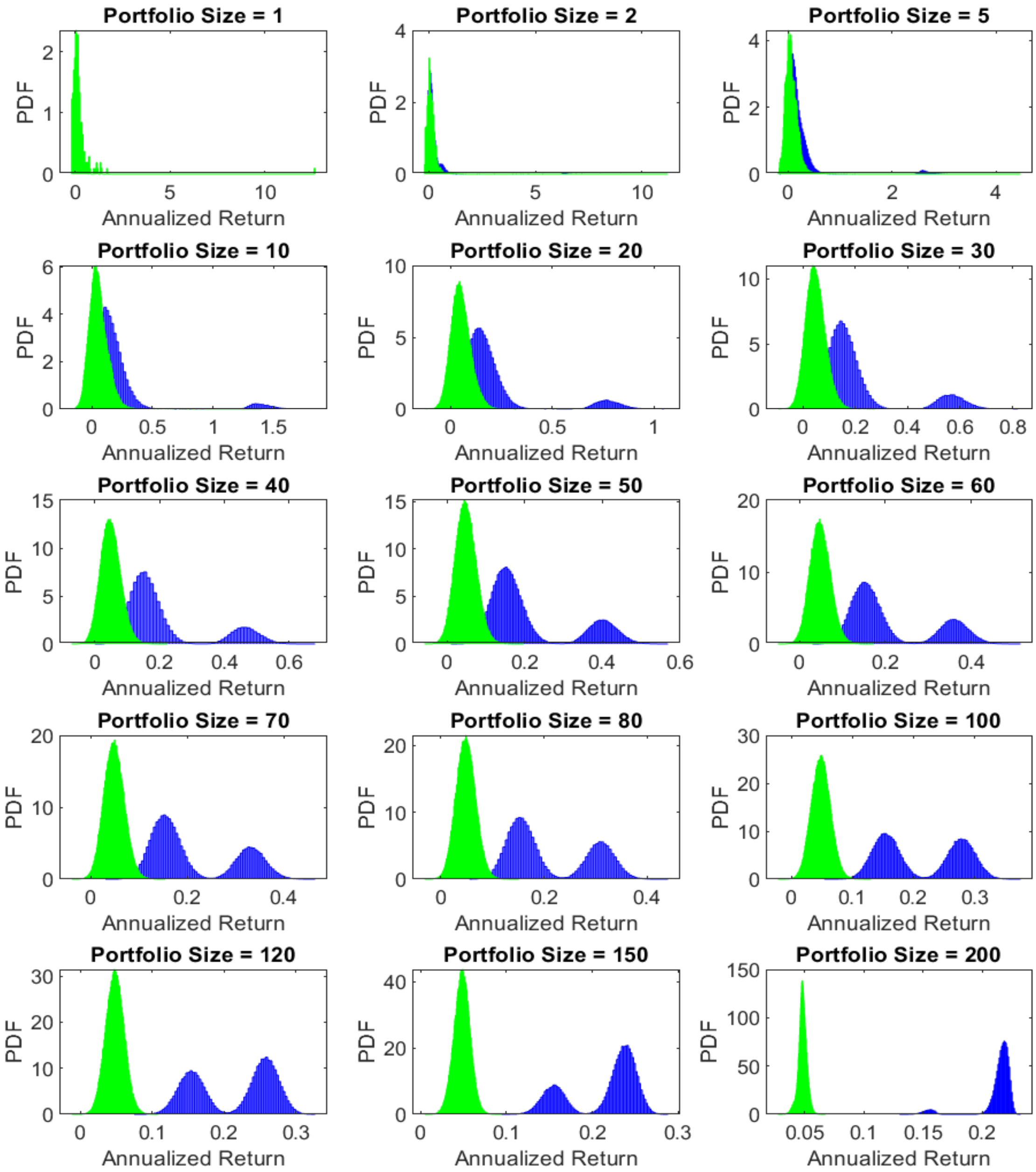

Figure 11: Return distributions for different portfolio sizes. This plot is based on 200.000 simulations between 2011-01-03 and 2015-12-30. Blue distribution is for equally weighted portfolios and green distribution is for capitalization weighted portfolios.

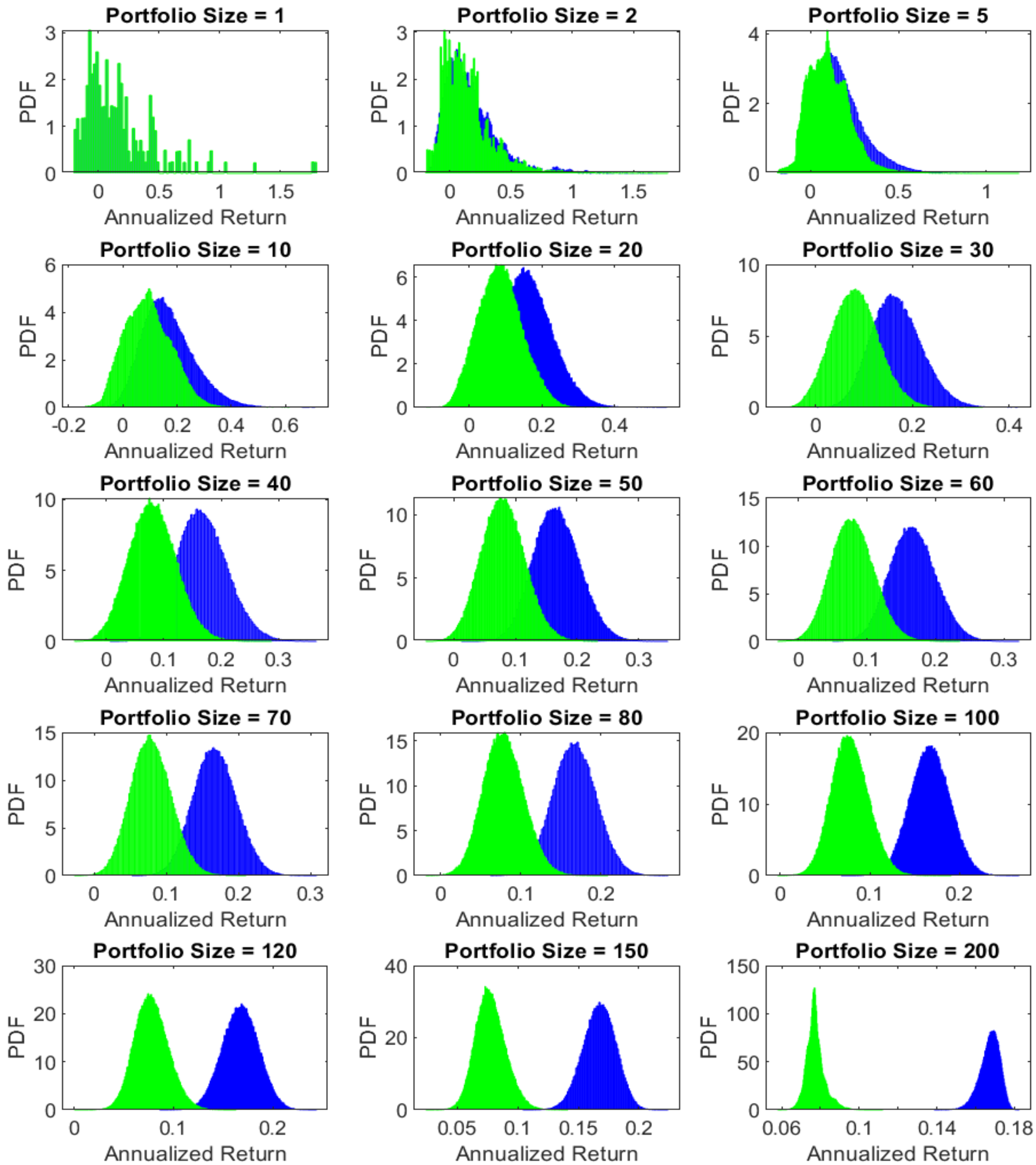

Figure 12: Return distributions for different portfolio sizes. This plot is based on 200.000 simulations between 2016-01-04 and 2020-12-30. Blue distribution is for equally weighted portfolios and green distribution is for capitalization weighted portfolios.

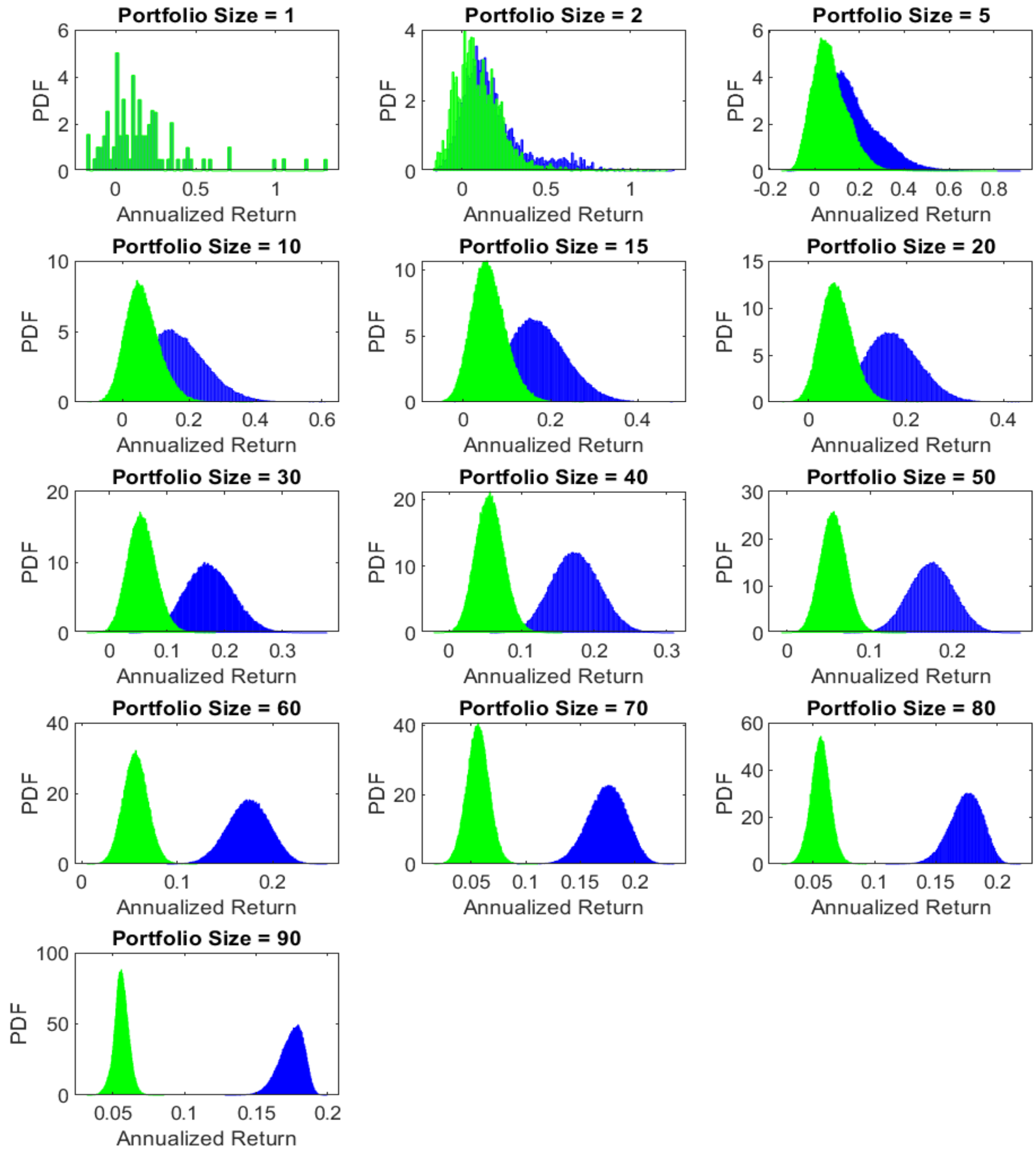

Figure 13: Return distributions for different portfolio sizes. This plot is based on 200.000 simulations between 2011-01-03 and 2015-12-30 and only containing Large Cap stocks. Blue distribution is for equally weighted portfolios and green distribution is for capitalization weighted portfolios. Notice, the bimodal property, seen in figure 11, is not visible here.

Appendix C

Populärvetenskaplig Sammanfattning (Swedish)

Varje aktieinvesterare måste, direkt eller indirekt, förhålla sig till portfölj diversifikation eller mer konkret frågan: *”Hur många aktier skall jag investera i?”*. Forskningens svar på denna frågan är att det varierar mellan 10 till 300 stycken. Detta kan sättas i kontrast mot det faktum att 41.79 % av svenska aktiesparare år 2020 endast ägde en aktie.

Den stora variationen inom forskningen kan förklars med hjälp av flera faktorer såsom att olika marknadsförutsättningar och att olika tidsperioder studerats. Den enskilt *viktigaste* faktorn är dock att olika metoder använts för att svara på frågan och är huvudförklaringen till den stora variationen.

Volatilitet eller standardavvikelse är ett viktigt risk begrepp inom finans. Det är ett mått på hur mycket din portfölj eller aktie rör sig från dag till dag (andra tidsspann såsom veckor eller månader kan också användas). En hög volatilitet medför att du kan förvänta dig stora svängningar. Studier har visat att det räcker med ca 10 aktier för att volatiliteten skall stabilisera sig och inte kunna minskas mycket mer. Men, konsensus bland forskningen verkar vara att detta risk mått, enskilt betraktat, inte beaktar viktiga aspekter vad gäller risk. En sådan aspekt är investerares grad av (o)säkerhet i att portföljavkastningen kommer skilja sig från en specifik samling av aktier, t.ex. en marknad såsom Large Cap aktier i Sverige. För att inkludera denna dimension av risk så kan konceptet Shortfall i % (Engelsk terminologi) användas. Måttet är definierat som skillnaden i avkastningen på en specifik samling av aktier (t.ex. alla Large Cap aktier) och 5% percentilen för en enskild aktieportfölj med ett fixt antal aktier dividerat med avkastningen på en specifik samling av aktier. Vill man använda någon annan percentil eller samling av aktier så är det upp till vardera investerare vilket gör måttet flexibelt.

Slutsatsen av att använda Shortfall i % blir att ca 120-150 aktier krävs för att uppnå en Shortfall på 20% under en 10 års period. För fem (ett) års perioder krävs 150 (70-80) aktier. I samtliga fall studerades lika viktade portföljer och samlingen av aktier som studerades var bolag från Small, Mid och Large Cap listan. Som privat sparare kan man enklast uppnå denna grad av diversifikation genom att investera i fonder. Enskilda aktier kan läggas till i portföljen enligt vardera investerares behag.

Det visades även att överlag så kommer investerare föredra lika viktade portföljer framför kapitaliserings viktade portföljer. Denna rangordning kunde göras med hjälp av ett teretiskt koncept som kallas för andra-gradens stokastisk dominans. Det är ett bra rangordnings verktyg då alla investere som ogillar risk och vill maximera sin nytta föredrar en portfölj som stokastiskt dominerar en annan. Anledningen till varför man använder det teoretiska konceptet stokastisk dominans beror på att rangordningen då kan göras på ett logiskt och transparent sätt.

I studien visades också att portföljer som hålls i 10 år stokastiskt dominerar portföljer som hålls under fem eller ett års perioder.

Går det att försona det faktum att 41.79 % av aktiespararna endast ägde en aktie? Som en proxy så var medianportföljens värde år 2016 30 000 kr. Skulle värsta scenariot hända att man förlorar alla pengar skulle det inte növädnigtvis innebära en kris för enskilda sparares privatekonomi samtidigt som det finns en chans att man tjänar mycket pengar genom att välja nästa "vinnar aktie". Därför kan beteendet förklars genom att folk helt enkelt "tar en chansning". Om man förvaltar en större förmögenhet skulle de allra flesta mest sannolikt föredra att vara mer diversifierade istället för att endast äga en aktie.

Slutligen bör det nämnas att diversifikation i praktiken är ett mångfacetterat begrepp, t.ex. kan man diversifiera över olika tillgångsslag såsom obligationer och råvaror. Man kan även diversifiera över tid. Denna uppsats beaktar inte dessa aspekter men det vore en intressant infallsvinkel för fortsatta studier.

References

Aktiespararna (2021). Kontrollera riskerna (in swedish). https://www.aktiespararna.se/utbildning/kontrollera-riskerna. Accessed: 2021-04-21.

Benjelloun, H. (2010). Evans and archer-forty years later. *Investment Management and Financial Innovations*, 7(1):98–104.

Bird, R. and Tippett, M. (1986). Note—naive diversification and portfolio risk—a note. *Management Science*, 32(2):244–251.

Black, F. and Litterman, R. (1992). Global portfolio optimization. *Financial analysts journal*, 48(5):28–43.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of Financial studies*, 22(5):1915–1953.

Domian, D. L., Louton, D. A., and Racine, M. D. (2003). Portfolio diversification for long holding periods: how many stocks do investors need? *Studies in Economics and Finance*.

Domian, D. L., Louton, D. A., and Racine, M. D. (2007). Diversification in portfolios of individual stocks: 100 stocks are not enough. *Financial Review*, 42(4):557–570.

Elton, E. J. and Gruber, M. J. (1977). Risk reduction and portfolio size: An analytical solution. *The Journal of Business*, 50(4):415–437.

Engle, R. F., Ledoit, O., and Wolf, M. (2019). Large dynamic covariance matrices. *Journal of Business & Economic Statistics*, 37(2):363–375.

Euroclear (2020). Aktieägandet i sverige 2020 (in swedish). `https://www.euroclear.com/sweden/sv/det-svenska-aktieagandet.html`. Accessed: 2021-05-04.

Evans, J. L. and Archer, S. H. (1968). Diversification and the reduction of dispersion: an empirical analysis. *The Journal of Finance*, 23(5):761–767.

Fisher, L. and Lorie, J. H. (1970). Some studies of variability of returns on investments in common stocks. *The Journal of Business*, 43(2):99–134.

Green, J., Hand, J. R., and Zhang, X. F. (2013). The supraview of return predictive signals. *Review of Accounting Studies*, 18(3):692–730.

Jorion, P. (1992). Portfolio optimization in practice. *Financial analysts journal*, 48(1):68–74.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621.

Ledoit, O. and Wolf, M. (2004). Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119.

Ledoit, O. and Wolf, M. (2017). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *The Review of Financial Studies*, 30(12):4349–4388.

Levy, H. (1979). Does diversification always pay? *Portfolio Theory*, 25:63–72.

Mao, J. C. (1970). Essentials of portfolio diversification strategy. *Journal of Finance*, pages 1109–1121.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

Markowitz, H. (1959). *Portfolio Selection, Efficient Diversification of Investments*. John Wiley and Sons, New York.

Markowitz, H. M. (1976). Markowitz revisited. *Financial Analysts Journal*, 32(5):47–52.

Michaud, R. O. (1989). The markowitz optimization enigma: Is 'optimized'optimal? *Financial analysts journal*, 45(1):31–42.

Michaud, R. O. and Michaud, R. O. (2008). *Efficient asset management: a practical guide to stock portfolio optimization and asset allocation*. Oxford University Press, New York, 2nd edition.

Nasdaq (2021). Historiska kurser (in swedish). `http://www.nasdaqomxnordic.com/aktier/historiskakurser`. Accessed: 2021-04-21.

Newbould, G. D. and Poon, P. S. (1993). The minimum number of stocks needed for diversification. *Financial Practice and Education*, 3(2):85–87.

Newbould, G. D. and Poon, P. S. (1996). Portfolio risk, portfolio performance, and the indvidual investor. *The Journal of Investing*, 5(2):72–78.

Petters, A. O. and Dong, X. (2016). *An introduction to mathematical finance with applications*. Springer.

Rui Ming Daryl, C. and Kai Jie Shawn, L. (2012). The attenuation of idiosyncratic risk under alternative portfolio weighting strategies: recent evidence from the uk equity market. *International Journal of Economics and Finance*, 4(11).

SCB (2016). Shareholder statistics. `https://www.scb.se/en/finding-statistics/statistics-by-subject-area/financial-markets/shareholding-statistics/shareholders-statistics/pong/statistical-news/shareholder-statistics-december-2016/`. Accessed: 2021-05-04.

SCB (2021). Average income (in swedish). `https://www.scb.se/hitta-statistik/sverige-i-siffror/utbildning-jobb-och-pengar/medelloner-i-sverige/`. Accessed: 2021-05-04.

Sharpe, W. F. (1963). A simplified model for portfolio analysis. *Management science*, 9(2):277–293.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442.

Statman, M. (1987). How many stocks make a diversified portfolio? *Journal of financial and quantitative analysis*, pages 353–363.

Statman, M. (2004). The diversification puzzle. *Financial Analysts Journal*, 60(4):44–53.

Upson, R. B., Jessup, P. F., and Matsumoto, K. (1975). Portfolio diversification strategies. *Financial Analysts Journal*, 31(3):86–88.

Whang, Y.-J. (2019). *Econometric analysis of stochastic dominance: concepts, methods, tools, and applications*. Cambridge University Press.

Wilmott, P. (2006). *Paul Willmott on Quantitiative Finance*. John Wiley Sons, Wiltshire, 2nd edition.

Zweig, J. (1998). America's top pension fund. *Money*, 27(1):114.

Master's Theses in Mathematical Sciences 2021:E20
ISSN 1404-6342

LUNFMS-3097-2021

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/