1. Introduction

1.1. **Prices, returns, holdings, and portfolios.** The goal of this course is to develop a quantitative approach to portfolio management. Such an approach is absolutely necessary when managing a systematic trading strategy, but the mathematical tools we will develop are useful for analyzing any portfolio, including portfolios built by fundamental managers.

Some fundamental analysts are more adept at analyzing companies than they are at building portfolios, and there are many aspects of portfolio construction that are simply too complex to be performed manually by humans. Hence ideas of fundamental managers are perhaps best monetized using mathematically sophisticated *portfolio construction* techniques like the ones we'll cover. All of the techniques we will discuss are actually used in most of the largest investment banks, asset management firms, and hedge funds in the world today.

Portfolio theory is defined broadly as the study of any set of holdings involving one or more assets. The assets involved could be financial assets, such as shares in corporations, or they could be non-financial real assets such as gold bars or land. An important subfield of portfolio theory is the study of optimization: how to form portfolios that are optimal with respect to the investor's utility function, but we will also develop tools to study portfolios which weren't necessarily constructed optimally.

Portfolio theory is itself a subfield or specialization of the more general topic of decision-making under uncertainty. Any decision we make has various possible outcomes, usually with some element of chance or luck involved, and with various levels of happiness or satisfaction associated to the various outcomes. When the *decision* is whether to hold your current portfolio, or trade into a different one, then we arrive at the special case of decision theory known as portfolio theory.

Changes in portfolio holdings are called *trades*. Although it wasn't emphasized in classical studies of portfolio theory, trading itself is a complex process with many decision variables that need to be considered and potentially optimized. The topic of market microstructure is a full-semester course on its own.

For now, suffice it to say that all trading is costly, and the costs from trading can be a meaningful part of the ultimate profitability of a given trading strategy or fund. Indeed, some putative "arbitrage opportunities" don't exist at all once costs are properly accounted for, and I'll wager that most quantitatively-oriented hedge funds have only a vague idea of their capacity (how much capital they can actually manage without unacceptable degradation in returns).

We must walk before we can run. Hence our very first forays into the world of portfolio management will be concerned with establishing a coherent notation

and a coherent language for describing the fundamental actors in our story: prices, returns, holdings, and portfolios.

Loosely, or in conversation, one sometimes speaks of "**the** price" of an asset, but in continuous limit order book markets (such as most of the world's equity markets in developed countries) there are always at least two prices.

Definition 1.1. A *bid price* is the highest price that a buyer is willing to pay for the security. The *ask price*, or offer, is the lowest price a seller will accept for the security. The *midpoint price* is the mathematical midpoint between the best bid and offer, $(p_{\text{bid}} + p_{\text{ask}})/2$. These prices are rarely the same: the ask is usually higher than the bid. If you are buying a stock, you pay the ask price. If you sell a stock, you receive the bid price. The difference between the two prices is called the *spread*.

In the absence of some form of predictability in the dynamics of the limit order book, a round-trip transaction of buying a share and selling it subsequently (or selling then buying to cover) should cost, in expectation, at least one bid-offer spread and possibly more.

In fact there are many kinds of "trading costs" and the full extent of these are sometimes only vaguely understood by practitioners. A humorous anecdote: a quantitative trader was once hired who, after a few months of trading, asked the operations group "what are financing costs?" The operations group dutifully went away, ran some numbers through spreadsheets, and came back with some number like $752,000. The trader responded, "no, I mean, what ARE financing costs – like what does the word mean?"

If you are trading equities, your trading costs likely include various highly-predictable or exactly calculable costs such as: borrow fees if one is selling short, financing costs covering the use of leverage on the long positions, transaction taxes in some markets, taxes on dividends you receive, commissions and small-order charges or ticket charges, and the bid-offer spread cost on aggressive orders already mentioned above.

Costs also include things you cannot know exactly (even after the fact!), but you can only build a model to try and predict, such as temporary and permanent price impact. (The manager of a trading group must also consider the costs of data, computers, and personnel.) One implication of all these costs is that strategies with high turnover and/or requiring leverage must also have *extremely* accurate forecasts.

Let all of the assets we are going to consider for our portfolio be indexed by $i = 1, \ldots, n$. They do not need to be stocks, but in some cases we will use equities as convenient examples, or use language derived from the equity markets.

Asset prices at time t will be denoted p_t with p_t^i the price for the i-th asset. The *return* of the i-th asset over the interval $[t, t+1]$ will be denoted

$$r_{t+1}^i = p_{t+1}^i / p_t^i - 1. \tag{1.1}$$

As mentioned above, there are various prices one can use in return calculations. One could use the midpoint price as in (1.1), or one could use the last trade price before some cutoff time (such as the market close time). One could also use the volume-weighted average price (VWAP) over some interval. Moreover, one could convert both of the prices in (1.1) to some other currency. There are hence many definitions of *return* appropriate for different purposes.

The prices p_t^i sometimes need to be adjusted in order that the returns (1.1) represent the total return from holding the asset. For example, in the equity markets, stock splits and dividends are common occurrences.

Definition 1.2. A *stock split* is an issue of new shares in a company to existing shareholders in proportion to their current holdings. A *dividend* is a sum of money paid by a company to its shareholders out of its profits or reserves.

Stock splits are planned and organized so they occur between when one trading session ends and the next session begins; you can imagine the confusion if a stock's price changed by a large multiple in the middle of the trading day! When a stock split occurs, the price instantaneously adjusts before the next trading session begins to reflect the dilution factor. A similar effect occurs for dividends on the ex-dividend date – the stock price adjusts somewhat.

If there are stock splits or dividends between t and $t+1$, the effect of those should be included in the return variable, for example by adjusting the price before the split to be in the same units as price after the split before calculating (1.1). Once this adjustment has been made, the quantity (1.1) is called *total return*. The following example data illustrates a 7-to-1 split for Apple Inc (AAPL), sourced from the CRSP daily stock file.

t1	p1	t2	p2	total.return
2014-06-04	644.82	2014-06-05	647.35	0.003924
2014-06-05	647.35	2014-06-06	645.57	-0.00275
2014-06-06	645.57	2014-06-09	93.7	0.016001
2014-06-09	93.7	2014-06-10	94.25	0.00587
2014-06-10	94.25	2014-06-11	93.86	-0.004138

In handling stock splits, care is required to avoid lookahead bias. If a stock price goes from 100 to 1000, the company is more likely to consider a split, because

higher share prices make the stock less accessible to individual retail investors. If one could look into the future and observe the set of stocks that were going to split, one would have a very profitable (fictitious) strategy. Hence the variable $1/p_t$, where p_t is adjusted to be in units of today's shares, is a (spurious, un-implementable) forecast. It isn't implementable because it has a *lookahead bias*, which is defined as some way of looking into the future in a backtest.

Fortunately, there is a foolproof way to avoid confusion and costly errors surrounding stock splits. It is the same as the way one often checks calculations in physics: make sure the units match up. The SI unit of force is a kilogram-meter per second-squared, so if you've done a force calculation and the result has dimensions of kilogram per second, you know you made a mistake. In finance and in physics, get into the habit of explicitly retaining the units on each variable that goes into a calculation.

A length measurement might have units of inches or centimeters, and you can convert from one to the other by applying the appropriate conversion factor. If a stock price p_t is displayed as 5 euros on a Bloomberg terminal on day t, this is, technically, an imprecise specification of units. A more complete specification is

$$\frac{\text{euros}(t)}{\text{shares}(t)}$$

where we read euros(t) as "euros on day t" and the denominator shares(t) is pronounced "shares on day t."

Note that "shares on day t" and "shares on day $t+1$" are *different* units. The currency (euros, in this case) is also part of the specification of units. In this notation, a split factor is simply the unit conversion factor shares(t)/shares$(t+1)$.

Suppose we want to compare a price on day t with a price on day $t+1$. A price specified in currency per "shares on day $t+1$" must be converted into units of "shares on day t" before any mathematical operation combining the two prices (such as division, to compute the return) could meaningfully be attempted. A sophisticated programmer could structure code in such a way as to render such meaningless comparisons impossible.

When doing financial calculations for corporate valuation purposes, units are even more important. Some measures, such as revenues, are typically specified as whole-company measures (and may be expressed in thousands on a financial statement) and multinational companies have revenues in various currencies, while other measures, such as earnings, are often reported and forecasted on a per-share basis. If an analyst forecasted EPS for Apple on 2014-06-06, was it in units of per-June6-shares or per-June7-shares?

We will also sometimes work with *log-relative returns* or "logrels" for short, which are

$$\log(1 + r^i_{t+1}) = \log p^i_{t+1} - \log p^i_t.$$

The *compound return* over many periods is defined as

$$R_T := \prod_{t=1}^{T}(1 + r^i_t) - 1$$

The logrel version of R_T is potentially easier to deal with:

$$\log(1 + R_T) := \sum_{t=1}^{T} \log(1 + r^i_t)$$

In other words, adding logrels corresponds to compounding.

Logrels are something it makes sense to add, while returns can be added only in an approximate calculation. For small r, one has $\log(1 + r) \approx r$ using the Taylor series, so we will sometimes model returns as linear functions of other variables.

Definition 1.3. Boldface notation without a superscript, $\boldsymbol{r}_{t+1}$, denotes the entire $n \times 1$ column vector of asset returns, typically referred to as a "cross-section" of asset returns.

When $t + 1$ is in the future, $\boldsymbol{r}_{t+1}$ is a random variable with values in $\mathbb{R}^n$. We do not know the probability density function (p.d.f.) of this random variable. Indeed, much of the field of empirical finance is dedicated to developing methods which can help to discover information, knowable at time t, that is relevant to the multivariate density of $\boldsymbol{r}_{t+1}$. Any structure inherent in the random process driving $\boldsymbol{r}_{t+1}$ potentially gives rise to analogous structure in the portfolio's return.

Definition 1.4. Throughout this course, we will let h^i_t denote our holdings in the i-th security at time t. (Mnemonic: h is the first letter in *holdings*.) The number h^i_t will always have units either of US dollars, or another appropriate currency converted to dollars. The portfolio at time t is $\boldsymbol{h}_t \in \mathbb{R}^n$.

Note that a stock portfolio is, in actuality, a collection of holdings of *shares*, not dollar amounts as per the previous definition. To convert shares into the dollar values of the holdings, one must multiply by a price. Hence $h^i_t = s^i_t p^i_t$. As the prices change, so do the values of h^i_t even in the absence of any trading. The change in the price is equivalently expressed as a return as per (1.1).

Then $\boldsymbol{h}_t \cdot \boldsymbol{r}_{t+1}$ is the portfolio's (unrealized) profit and loss (abbreviated P&L) over the interval $[t, t + 1]$, where the dot denotes a scalar product or "dot product" of the two vectors. Information (eg. from a statistical model's forecast) about the

variance-covariance matrix $\mathrm{cov}(\boldsymbol{r}_{t+1})$ potentially gives information about $\sigma^2(\boldsymbol{h}_t)$, the portfolio variance:

$$\sigma^2(\boldsymbol{h}_t) := \mathbb{V}[\boldsymbol{h}_t \cdot \boldsymbol{r}_{t+1}] = \boldsymbol{h}_t' \,\mathrm{cov}(\boldsymbol{r}_{t+1})\boldsymbol{h}_t \,.$$

Minimum volatility funds seek to minimize $\sigma^2(\boldsymbol{h}_t)$ while remaining fully invested. This is harder than it sounds, and we will discuss this problem in some detail later.

The vector $\boldsymbol{r}_{t+1}^{\mathrm{rf}}$ denotes the relevant vector of *risk-free rates*. If the asset returns one is modeling are local currency returns, one must be careful to have the correct risk-free rate for each currency!

Definition 1.5. *Excess returns* are the asset returns minus the appropriate risk-free rate of return:

$$\underset{[n\times 1]}{\boldsymbol{r}_{t+1} - \boldsymbol{r}_{t+1}^{\mathrm{rf}}}$$

where for concreteness, we will often indicate the matrix dimensions of each matrix or vector below it.

Many classical results in finance are derived under simplistic assumptions, which we know aren't exactly satisfied in reality. The most common such assumption is normality; asset returns are *never* normally distributed, and yet, it's probably the most common of the various assumptions in finance that are made to aid in computational tractability. A normal distribution is entirely specified if you know its first two moments (mean vector and covariance matrix).

Many sources I've come across, including published articles in prestigious academic journals, are confused about normal distributions and when normality is needed as an assumption. A very common misconception is that mean-variance optimization assumes normality of returns – it doesn't! Neither does the capital asset pricing model.

It makes sense to consider returns above the relevant risk-free rate when evaluating risky investments. A risky investment with expected return below or equal that of the risk-free investment would be strictly dominated by the risk-free investment. It is for this reason that the Sharpe ratio, defined below, is defined using excess returns:

Definition 1.6. In 1966, William F. Sharpe developed the *Sharpe ratio*:

$$S = \frac{\mathbb{E}[r - r_f]}{\sqrt{\text{var}[r]}}.$$

where r is a scalar-valued random variable denoting the return of an investment. The universal convention is that, whatever frequency the returns are sampled at (daily, weekly, monthly, ...), Sharpe ratio calculations always use annualized returns and annualized volatility.

Here $\mathbb{E}[\,]$ denotes the expectation value of a random variable. As such it does not directly refer to past returns; it is an aspect of a model or a forecast of the future. To emphasize this, the phrase *ex ante* (from Latin ex 'from, out of' + ante 'before.') is often used. One could, of course, also compute the statistic S on past (realized) returns, giving an *ex post* or *realized* Sharpe ratio. The Sharpe ratio is closely related to the Student's t-test for the null hypothesis of whether the true mean is zero.

In testing the null hypothesis that the population mean is equal to zero, one uses the Student t-test, based on the statistic $t = \overline{x}/(s/\sqrt{n})$ where $\overline{x}$ is the sample mean, s is the sample standard deviation, and n is the sample size. If the sample is of daily returns, and if $n = 252$ (the number of trading days in a year), then this equals the Sharpe ratio, but as the sample size grows, the t-statistic also increases (we are increasingly sure that the mean isn't zero), but the Sharpe ratio simply converges to its asymptotic value.

When applied to the returns of a mutual fund, the Sharpe ratio is not the ultimate measure of a manager's skill. This is because most mutual funds are constrained in some way which entails that they cannot simply seek absolute return over the risk-free rate. For example, it is almost impossible for a sufficiently diversified long-only fund to realize a positive Sharpe ratio over a period when the market is down substantially, no matter how great the manager's skill.

Definition 1.7. A *benchmark* is a standard against which the performance of a security, mutual fund or investment manager can be measured. Usually, broad market and market-segment stock and bond indexes are used for this purpose.

In a year like 2008, it's almost impossible for a well-diversified long-only US equity portfolio to have done well in absolute terms, regardless of the skill of the manager. Hence a manager's skill is closely related to performance *relative to the appropriate benchmark*. For this reason, one often considers the information ratio (IR), which is the same formula as the Sharpe ratio, but with the risk-free rate being replaced by an appropriate benchmark. Some managers do in fact have the

risk-free rate as their benchmark; these are so-called *absolute return* strategies, but this is more typical for hedge funds than for mutual funds.

1.2. **Lotteries and Rewards.** We need some precise mathematical definitions to formalize a concept that you already understand: a lottery. Let $\mathcal{X}$ denote a space of possible prizes (or "rewards"). For example, consider a game in which you flip a coin and win one dollar if the result is heads; if it's tails, you don't win anything. The space of possible rewards is then $\{0, 1\}$. A *lottery* is a probability distribution over the reward space. If we play the above game with a fair coin, this represents a different lottery than if we play it with a weighted or two-headed coin.

First suppose the reward space is finite; $\mathcal{X} = (x_1, \ldots, x_n)$. In this case, a lottery is simply a set of probabilities $\boldsymbol{p} = (p_1, \ldots, p_n)$ such that

$$p_i \geq 0 \ \forall i = 1, \ldots, n \ \text{ and } \ \sum_{i=1}^{n} p_i = 1.$$

Let $\mathcal{P} = \Delta(\mathcal{X})$ denote the space of lotteries, ie. discrete probability distributions. If we think of $\boldsymbol{p} = (p_1, \ldots, p_n)$ as a vector in $\mathbb{R}^n$, then geometrically $\Delta(\mathcal{X})$ is a simplex, and one reads $\Delta(\mathcal{X})$ as "the simplex over $\mathcal{X}$" .

This generalizes easily to the case where $\mathcal{X}$ is countably infinite. In that case, the probabilities must satisfy $\sum_{i=1}^{\infty} p_i = 1$ and the process of visualizing $\Delta(\mathcal{X})$ as a simplex takes quite a bit more imagination.

Finally, the mathematics of lotteries can be generalized to the case when $\mathcal{X}$ is a continuous space such as $\mathbb{R}$. A lottery then becomes a probability measure. Ownership of a stock (or more generally, any risky asset including an option) is an example of a lottery where $\mathcal{X}$ is a continuous space.

The first really important question about lotteries is whether two investors presented with the same lottery should value it in the same way. Swiss mathematician Daniel Bernoulli, in 1738, changed the way the world would view this question forever. Famous scientists before Bernoulli (among them Pascal and Fermat) had argued that the value of a lottery should be equal to its mathematical expectation and hence identical for all people, independently of their risk attitude. We shall see very shortly that this cannot be correct.

We now embark on a study of the classic theory of decision-making under uncertainty, which is fundamental to much of modern-day economics and finance. The story essentially began when Nicolas Bernoulli submitted five problems to the mathematician Pierre Rémond de Montmort in 1713.

Nicolas Bernoulli's five problems are reproduced as an appendix to the second edition of Montmort, 1714 "L'analyse sur les jeux de hazard", p. 402. The first satisfactory resolution of these problems was achieved about 20 years later when Nicolas Bernoulli's cousin, Daniel Bernoulli, wrote a Latin manuscript (translated

in Bernoulli (1954)) entitled: "Exposition of a New Theory on the Measurement of Risk."

The last of these problems, known as the "St. Petersburg paradox" in honor of Daniel Bernoulli's paper having appeared in *Papers of the Imperial Academy of Sciences in Petersburg*, runs as follows (see Montmort (1714) for the original statement). Assume that we agree to flip a fair coin, stopping at the first tails, let n be the number of flips, and you win 2^n dollars. The worst case scenario is $n = 1$ (tail on the first flip), so you always win at least 2 dollars and you would therefore pay a positive amount to play, but how much? The expectation value is

$$\frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 4 + \ldots = \sum_{n=1}^{\infty} \frac{1}{2^n} \cdot 2^n$$

Among other things, Bernoulli (1954) contains a proposed resolution to the paradox: if investors have a logarithmic utility, then their expected change in utility of wealth by playing the game is finite.

Of course, if our only goal were to study the St. Petersburg paradox, then there are other, more practical resolutions. In the St. Petersburg lottery only very unlikely events yield the high prizes that lead to an infinite expected value, so the expected value becomes finite if we are willing to, as a practical matter, disregard events which are expected to occur less than once in the entire lifetime of the universe.

Moreover, the expected value of the lottery, even when played against a casino with the largest resources realistically conceivable, is quite modest. If the total resources (or total maximum jackpot) of the casino are W dollars, then $L = \lfloor \log_2(W) \rfloor$ is the maximum number of times the casino can play before it no longer fully covers the next bet. Hence the mutually exclusive outcomes are: for $k = 1, 2, \ldots, L$ you win 2^k with probability 2^{-k}, or with probability $1 - \sum_{k=1}^{L} 2^{-k}$ you win W (the full jackpot). But note that, simplifying,

$$1 - \sum_{k=1}^{L} 2^{-k} = 2^{-L}$$

so the expected value is $L + 2^{-L}W$. If the casino has $W =$ one billion dollars, $L = 9 \ln 10 / \ln 2 \approx 29.9$ and the expected value of the "realistic Petersburg lottery" is only about 31 dollars.

The celebrity of the St. Petersburg paradox has overshadowed the other two examples given by Daniel Bernoulli that show that, most of the time, the value of a lottery is not equal to its mathematical expectation. One of these two examples, which presents the case of an individual named "Sempronius," anticipates the central contributions that will be made to risk theory about 230 years later by Arrow, Pratt and others.

We explain the Sempronius example by means of a humorous story of adventure on the high seas. The year is 1776 and you own goods located abroad, worth the equivalent of 1 standard-size gold bar. These goods cannot increase your wealth until they are shipped back to you via sailing vessels on the high seas, but it's a perilous journey; the probability that a ship is lost at sea is 1/2.

probably okay

maybe not...

You were planning to have the entire load sent on one ship. Captain Cook advises you that this is unwise and generously offers to split the load in half and send each half on a separate ship at no extra cost. Should you accept Cook's offer?

Let time T denote some arbitrary future time after which the ships will either have arrived, or are known to be lost. Assume your current wealth is zero, and let w_T denote your future wealth at time T. Calculate your expected future wealth in either case:

$$\text{one ship: } \mathbb{E}[w_T] = \frac{1}{2} \times 1 \qquad\qquad = 0.5 \text{ gold bar}$$

$$\text{two ships: } \mathbb{E}[w_T] = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = 0.5 \text{ gold bar}$$

You are about to advise Captain Cook that, due to extremely clever use of probability theory, you have proven that it doesn't matter – he can simply use one ship. Just then, Professor Daniel Bernoulli arrives and advises you to instead calculate $\mathbb{E}[u(w_T)]$ where $u(w) := 1 - e^{-w}$, leading to:

$$\begin{aligned}
\text{one ship: } \mathbb{E}[1 - e^{-w_T}] &= \frac{1}{2} \times (1 - e^{-1}) \\
&\approx 0.32 \\
\text{two ships: } \mathbb{E}[1 - e^{-w_T}] &= \frac{1}{4}(1 - e^{-0}) \\
&\quad + \frac{1}{2}(1 - e^{-1/2}) + \frac{1}{4}(1 - e^{-1}) \\
&\approx 0.35
\end{aligned}$$

Using Bernoulli's method, it seems that two ships are preferred, although the reason for the method's efficacy is perhaps still obscure. Bernoulli asks if you bothered to consider the *risk* when you compared the two scenarios. You reply angrily that you prefer to act first, and consider the risks later. But to make Bernoulli happy, you calculate:

$$\begin{aligned} \text{one ship:} \quad \mathbb{V}[w_T] &= \frac{1}{2}(0-0.5)^2 + \frac{1}{2}(1-0.5)^2 \\ &= 0.25 \\ \text{two ships:} \quad \mathbb{V}[w_T] &= \frac{1}{4}(0-0.5)^2 + \frac{1}{2}(0.5-0.5)^2 + \frac{1}{4}(1-0.5)^2 \\ &= 0.125 \end{aligned}$$

We then begin to see the benefits of the two-ship method. For the same expected return, it has lower risk. Bernoulli's funny-looking concave function of wealth has somehow captured that.

Bernoulli further explains that if

$$u(w) = \frac{1-\exp(-\kappa w)}{\kappa}$$

where $\kappa > 0$ is some positive scalar, then supposing w_T is normal,

$$\mathbb{E}[u(w_T)] = u\left(\mathbb{E}[w_T] - \frac{\kappa}{2}\mathbb{V}[w_T]\right) \tag{1.2}$$

This implies that maximizing $\mathbb{E}[u(w_T)]$ is equivalent to maximizing

$$\mathbb{E}[w_T] - \frac{\kappa}{2}\mathbb{V}[w_T]$$

since u is monotone. It turns out this is true for many fat-tailed distributions as well.

To recap, given two lotteries with the same expected reward, most people would prefer the one with lower risk. With a suitably appropriate choice of function $u()$, if we compare lotteries according to $\mathbb{E}[u(w_T)]$ then we will indeed choose the one with lower risk.

1.3. **Utility.** Bernoulli suggests that a lottery should be valued according to the "expected utility" that it provides. Instead of computing the expectation of the monetary outcomes, we should use the expected utility of the wealth. This relationship is characterized by a utility function u, which for every wealth level x tells us the level of "satisfaction" or "utility" attained by the agent with this wealth.

Although the function u transforms the objective result x into a perception, this transformation is assumed to exhibit some basic properties of rational behavior. For example, the function should be increasing in x. Bernoulli argues that if the utility u is not only increasing but also concave, then the two-ship version of the story will have higher expected utility than the one-ship version in accordance with

intuition. The concavity of the relationship between wealth x and utility u is quite natural. It implies that the marginal utility of wealth is decreasing with wealth: one values a one-dollar increase in wealth more when one is poorer than when one is richer.

Definition 1.8. An agent is *risk-averse* if, at any wealth level w, she dislikes every lottery with an expected payoff of zero: $\forall\, w, \forall\, \tilde{z}$ with $E[\tilde{z}] = 0$,

$$E[u(w + \tilde{z})] \leq u(w).$$

Any lottery $\tilde{z}$ with non-zero expected payoff can be decomposed into its expected payoff $E[\tilde{z}]$ and a zero-mean lottery $\tilde{z} - E[\tilde{z}]$. Thus a risk-averse agent always prefers receiving the expected outcome of a lottery with certainty, rather than the lottery itself. This implies that, for any lottery $\tilde{z}$ and for any initial wealth w,

$$E[u(w + \tilde{z})] \leq u(w + E[\tilde{z}])$$

相比于lottery，人们更倾向于现金.

Proposition 1.1. An agent is risk-averse, ie. inequality

$$E[u(w + \tilde{z})] \leq u(w + E[\tilde{z}])$$

holds for all w and $\tilde{z}$, if and only if $u()$ is concave.

The above Proposition is a rewriting of the famous Jensen's inequality. Consider any real-valued function ϕ. Jensen's inequality states that $E\phi(\tilde{y}) \leq \phi(E[\tilde{y}])$ for any random variable $\tilde{y}$ if and only if ϕ is concave.

One way to measure the degree of risk aversion of an agent is to ask how much she is ready to pay to get rid of a zero mean risk $\tilde{z}$. The answer to this question will be referred to as the risk premium Π associated to that risk. Mathematically, the risk premium is defined by

$$E[u(w + \tilde{z})] = u(w - \Pi)$$

One very convenient property of the risk premium, is that it is measured in the same units as wealth. This allows us to define how much wealth a given risk is worth.

Definition 1.9. The *certainty equivalent* e of risk $\tilde{z}$ is the sure increase in wealth that has the same effect on welfare as having to bear risk $\tilde{z}$, i.e.,

$$E[u(w + \tilde{z})] = u(w + e)$$

When $E[\tilde{z}] = 0$ one has $e = -\Pi$ which makes sense given the interpretation of Π as the amount one would pay to remove (with certainty!) the risk. A direct consequence of Proposition 1.1 is that the risk premium is nonnegative when u is concave, i.e. when the agent is risk-averse.

The risk premium is a complicated nonlinear function of the distribution of $\tilde{z}$, of initial wealth w and of the utility function u. We can attempt to approximate the risk premium using the Taylor series for $u()$:

$$\begin{aligned} u(w - \Pi) &\approx u(w) - \Pi u'(w) \\ E[u(w + \tilde{z})] &\approx u(w) + \frac{1}{2}\sigma^2 u''(w) \end{aligned}$$

Combining these, one has

$$\Pi \approx \frac{1}{2}\sigma^2 A(w) \quad \text{where} \quad A(w) := -\frac{u''(w)}{u'(w)} \tag{1.3}$$

Under risk aversion, function A is positive. It would be zero or negative respectively for a risk-neutral or risk-loving agent. $A(\cdot)$ is hereafter referred to as the *degree of absolute risk aversion* of the agent.

The approximation $\Pi \approx \frac{1}{2}\sigma^2 A(w)$ is known as the *Arrow-Pratt approximation*, as it was developed independently by Arrow (1963) and Pratt (1964). The cost of risk, as measured by the risk premium, is approximately proportional to the variance of its payoffs. Thus, the variance might appear to be a good measure of the degree of riskiness of a lottery. Kenneth Arrow was Professor of Economics at Harvard and the youngest person ever to receive the Nobel Prize in Economics.

A usefully simple set of examples comes from looking at when the Arrow–Pratt absolute risk aversion is constant:

$$A(w) := -\frac{u''(w)}{u'(w)} = \kappa$$

where $\kappa > 0$ is some arbitrary constant. The solutions are the *constant absolute risk-aversion (CARA)* utility functions, which are exponential functions characterized by

$$u(w) = -\frac{\exp(-\kappa w)}{\kappa}$$

where $\kappa > 0$ is some positive scalar. The distinguishing feature of CARA utility functions is constant absolute risk aversion: $A(w) = \kappa$ for all w.

Definition 1.10. An agent v is said to be *more risk-averse* than another agent u with the same initial wealth if the risk premium of any risk is larger for agent v than for agent u.

Proposition 1.2. The following three conditions are equivalent:
(a) Agent v is more risk-averse than agent u,
(b) For all w, $A_v(w) \geq A_u(w)$.
(c) Function v is a concave transformation of function u, meaning:

$$\exists \phi \quad \text{with } \phi' > 0 \quad \text{and} \quad \phi'' < 0 \quad \text{such that} \quad v(w) = \phi(u(w))$$

理解(c)：同样的wealth.
对于V来讲时显utility更低.
因为V是两个concave的
叠加 V(·)=φ(U(·))

A particularly easy special case to analyze is when the wealth random variable $\tilde{w}$ is normally distributed. In this case the Arrow-Pratt "approximation" is exact. Indeed, supposing $\tilde{w}$ has mean μ and variance σ^2, then

$$E[u(\tilde{w})] = u\left(\mu - \frac{1}{2}\kappa\sigma^2\right) \tag{1.4}$$

The relation (1.4) means that maximizing $E[u(\tilde{w})]$ is equivalent to maximizing

$$\mu - \frac{1}{2}\kappa\sigma^2,$$

since u is monotone.

Note that constant *absolute* risk aversion is a reasonable preference for a hedge fund who is optimizing a portfolio over the next few days and does not anticipate any large capital changes (such as subscriptions or redemptions) over the same period of time. Hence over the range of possible values of w that we're talking about, they'll have roughly the same aversion to (say) one million dollars of volatility.

Note – in the preceding argument, we have not assumed the actual utility function is quadratic. Quadratic functions don't make sense as utility functions, since it implies that beyond some level more wealth is somehow worse.

We henceforth assume that volatility in the wealth random variable $\tilde{w}$ comes from trading financial assets (as opposed to the stormy seas our friend Sempronius had to worry about!)

Let $\boldsymbol{h}_t \in \mathbb{R}^n$ denote the portfolio holdings, measured in dollars or an appropriate numeraire currency, at some time t in the future. Note that holdings can be positive or negative, where negative holdings correspond to short positions. Let $\boldsymbol{h}_0$ denote the current portfolio.

Let $\boldsymbol{r}_{t+1} \in \mathbb{R}^n$ denote the return over the interval $[t, t+1]$. Hence $\boldsymbol{r}_{t+1} \in \mathbb{R}^n$ is an n-dimensional vector whose i-th component is

$$r_{t+1,i} = p_i(t+1)/p_i(t) - 1$$

where $p_i(t)$ is the i-th asset's price at time t (adjusted for splits or capital actions if necessary).

Let w_0 denote the investor's current wealth at time $t = 0$, and let w_t denote the investor's total wealth in t periods. For all $t > 0$, w_t concerns future observations, and is therefore a random variable. Clearly the total wealth is a sum of the initial wealth, plus all increments (which can be positive or negative):

$$w_T = w_0 + \sum_{t=1}^{T} \delta w_t$$

For simplicity suppose we are only looking one period in the future, so $T = 1$. When considering single-period problems we can omit the various time subscripts, to keep notation simple. The expected-utility maximizer chooses $\boldsymbol{h}$ to satisfy

$$\boldsymbol{h}^* = \operatorname{argmax} \mathbb{E}[u(w_1)] = \operatorname{argmax} \mathbb{E}[u(w_0 + \boldsymbol{h}'\boldsymbol{r})] \tag{1.5}$$

If we're willing to make the Arrow–Pratt approximation (or we're working in a case where that approximation is exact), then (1.5) becomes a simpler problem:

$$\text{Maximize over } \boldsymbol{h}: \quad \mathbb{E}[\boldsymbol{h}'\boldsymbol{r}] - \frac{\kappa}{2}\mathbb{V}[\boldsymbol{h}'\boldsymbol{r}]$$

which can be further simplified to

$$\text{Maximize over } \boldsymbol{h}: \quad \boldsymbol{h}'\mathbb{E}[\boldsymbol{r}] - \frac{\kappa}{2}\boldsymbol{h}'\boldsymbol{\Omega}\boldsymbol{h}, \quad \boldsymbol{\Omega} := \operatorname{cov}(\boldsymbol{r}) \tag{1.6}$$

We have thus derived the Markowitz (1952) problem from the theory of decision-making under uncertainty. This is not how Markowitz derived the problem in his 1951 PhD thesis, but with the benefit of hindsight, Markowitz' theory fits very nicely into the modern understanding of decision-making with uncertainty.

Henceforth we assume that no asset is a perfect linear combination of the others. Actually maximizing $\boldsymbol{h}'\mathbb{E}[\boldsymbol{r}] - \frac{\kappa}{2}\boldsymbol{h}'\boldsymbol{\Omega}\boldsymbol{h}$ is more subtle than one might think. If $\boldsymbol{\Omega}$ admits zero as an eigenvalue, the solution is not even well-defined, but such zero eigenvalues are also nonsensical from a finance perspective. If zero is a non-trivial eigenvalue, and if v is the corresponding eigenvector, then we can think of v as proportional to portfolio weights, and such portfolio would have zero variance: $v'\boldsymbol{\Omega}v = 0$. If the assets are truly different, this is impossible. So $\boldsymbol{\Omega}$ must live in S^n_{++} and not just S^n_+.

If we rule out zero eigenvalues, then the vector $\boldsymbol{h}$ that solves

$$\text{Maximize over } \boldsymbol{h}: \quad \boldsymbol{h}'\mathbb{E}[\boldsymbol{r}] - \frac{\kappa}{2}\boldsymbol{h}'\boldsymbol{\Omega}\boldsymbol{h}, \quad \boldsymbol{\Omega} := \operatorname{cov}(\boldsymbol{r})$$

is given by

$$\boldsymbol{h}^* = (\kappa\boldsymbol{\Omega})^{-1}\mathbb{E}[\boldsymbol{r}]$$

For US equities $n \approx 1500$ to 3000 depending on whether our strategy includes small caps. In this range, estimating $\boldsymbol{\Omega}$ directly using sample covariances of the equity returns is pure lunacy. (This is obvious from the dimension of the parameter

space, not a profound or controversial statement.) With T historical periods, one has nT data points and $n(n+1)/2$ free parameters in the full covariance matrix, so $2T/(n+1)$ data points per parameter. If $n = 2500$ we need 10 years' history to get 2 data points per parameter! Let B be a $T \times n$ matrix having the stock return time series as columns. De-mean each column. Then the covariance matrix is $\mathbf{\Omega} \propto B'B \in S_+^n$. The rank of $B'B$ is at most T, so if $T < n$ it is impossible that $\mathbf{\Omega}$ is invertible.

Problem 1.1. Prove that an agent is risk-averse (definition 1.8) if and only if $u()$ is concave.

Problem 1.2. Consider an agent whose initial wealth is w, a constant, and final wealth is $\tilde{w} = w + \tilde{z}$, where $\tilde{z}$ is a lottery. Show that when $u(w) = -\exp(-\kappa w)/\kappa$ and $\tilde{w}$ is normally distributed with mean μ and variance σ^2, then the Arrow-Pratt approximation is exact.

Problem 1.3. Prove Proposition 1.2.

Problem 1.4. Consider a function $v(.)$ such that $v(x) = a + bu(x)$ for all x, for some pair of scalars a and b, where $b > 0$. Show that a decision-maker with utility function $v()$ makes the same decisions and has the same certainty-equivalents as a decision maker with utility function $u()$.

References

Arrow, Kenneth J (1963). "Liquidity preference, Lecture VI in "Lecture Notes for Economics 285, The Economics of Uncertainty", pp 33-53". In:

Bernoulli, Daniel (1954). "Exposition of a new theory on the measurement of risk". In: *Econometrica: Journal of the Econometric Society*, pp. 23–36.

Markowitz, Harry (1952). "Portfolio selection*". In: *The Journal of Finance* 7.1, pp. 77–91.

Montmort, Pierre Rémond de (1714). *Essai d'analyse sur les jeux de hazards. 2.* Jombert.

Pratt, John W (1964). "Risk aversion in the small and in the large". In: *Econometrica: Journal of the Econometric Society*, pp. 122–136.