A Sketch of Basic Probability and Statistics for Finance

Jon Faust

180.266
Johns Hopkins University
March 28, 2017

1 Intro and motivation

The following is a mixture of basic probability and statistics from a finance perspective and some practical comments about finance applications.

Formal statistical tools are at the center of an immense portion of finance and of many other practical fields. In formal statistical analysis in finance, we treat returns as random and apply conventional statistical tools to analyze them.

One of the great challenges in managing risk of any sort is simply finding a clear and accurate way to summarize the randomness one faces. Managers, corporate boards, and regulators rely heavily on standard statistical tools to summarize the range of likely outcomes. The narrowest goal of these notes is to provide working definitions for several relevant concepts: mean, expectation, variance, covariance, standard deviation, correlation, quantile, median, mode, interquartile range, and value at risk. These are all ways to summarize randomness.

The question of which summary statistics are most revealing of the relevant information is one of great practical importance. As consumers of statistical information, managers, boards and regulators need to know what each concept reveals and what it may tend to obscure. Similarly, individuals digesting 'raw' data for bosses and other stakeholders need to understand the subtleties of these measures: what will best inform? What might be TMI and what might be too little?

Many wise folks see breakdowns at every level here playing a role in the recent financial crisis. One goal of these notes is to help you begin building some perspective on the practical uses and abuses of statistical concepts. I hope you all will leave here less likely to blow up your personal savings, your companies, and the economy than earlier generations.

2 Foramlly characterizing randomness

A random variable is a variable the value of which is unknown, but will be determined according to some well-defined rules. Example: Before flipping a coin, whether the toss will come up heads or tails is a random variable. The value is ultimately resolved by the process of flipping and if this is a fair flip, the two possible outcomes are equally likely.

In formal analysis in finance, we treat risky returns as random variables.

To formally analyze any random variable, we must know i) an exhaustive list of all the values it may take on and ii) the probability that each outcome will be the one realized.

For example: if x describes the roll of a fair six-sided die, the possible outcomes are 1,. . . ,6; each has probability 1/6.

Note that since the list of outcomes is exhaustive, the probabilities must sum to one. Put another way, with probability 1, one of the outcomes happens.

In the coin toss and dice examples, the possible outcomes are a set of distinct values—$\{\text{heads}, \text{tails}\}$ or $\{1, \ldots, 6\}$. The alternative to these cases with a discrete set of outcomes are cases in which the possible outcomes are a range of real numbers.[1] For essentially all practical statistical work, we can limit consideration to a discrete set of outcomes. If you can do the statistical work using standard computers, the discrete case must be enough: everything on the computers is discrete![2]

Sticking with the discrete case means that a full characterization of the randomness can be expressed as a table in which each row gives a possible outcome and the probability with which that outcome will occur. For a die:

Table 1: Prob. distribution for fair die

pr	outcome
1/6	1
1/6	2
1/6	3
1/6	4
1/6	5
1/6	6

[1] In principle, if I spin the pointer on a dial labelled like the face of a clock, the result could be any real number between 0 and 12.

[2] More technically, the distinction here is between random variables with a countable set of outcomes versus an uncountable set.

You can always think of any random case and of any formal risk model as a table in which each row gives an outcome and the probability that the outcome will occur.

2.1 Why we need summary measures

In practical cases in many areas, we may need to take account of thousands or millions of possible outcomes. That is, our table has millions of rows. Further, each outcome might may itself have many characteristics.

For example, suppose I am modelling the value of my porfolio as of tomorrow. I own a portfolio of, say 50 stocks. An outcome in this case is a list of prices of each of the 50 stocks. And of course there are millions of possibilities for what the prices of the 50 stocks may be tomorrow. So we can think of my risk model as a table with 51 columns and millions of lines. To save paper, I'll jsut show a small of a hypothetical risk model:

Table 2: A Portfolio risk model

	Price ($) tomorrow of			
pr	IBM	Merk	...	Apple
0.0000013	150.23	64.57	...	120.61
⋮	⋮	⋮	⋮	⋮
0.0000002	205.45	20.25	...	117.56

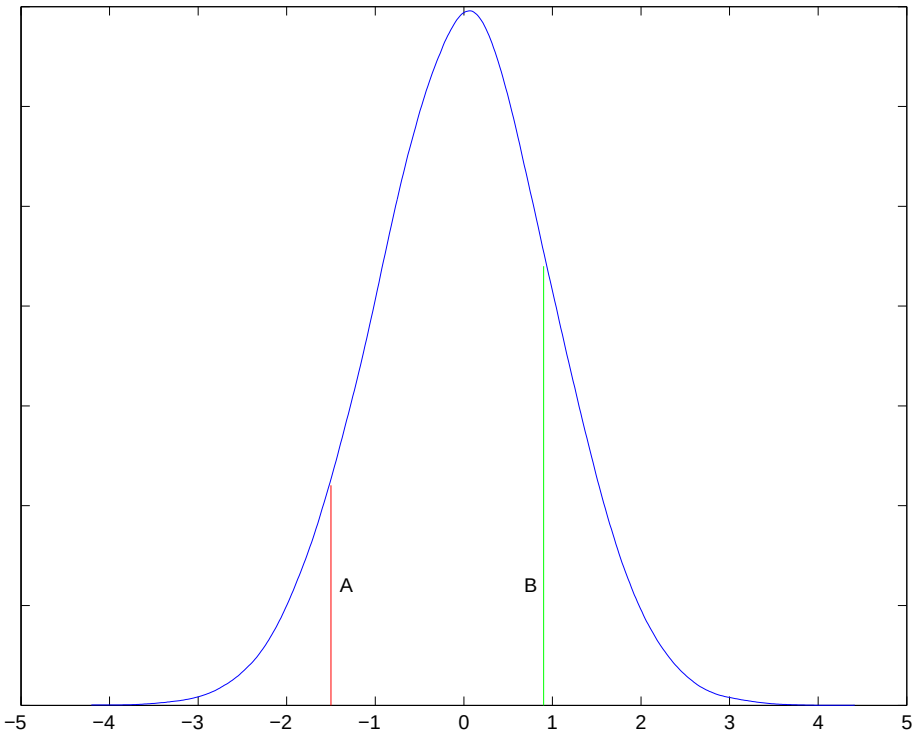
If you expand all those vertical and horizontal ellipses in your mind, you should have the image of very large set of numbers—so large as to be useless for an practical discussion unless it is summarized to pull out the most important features from the standpoint of some practical decision such as how much of my wealth should I put in each of the 50 stocks.

So the trick in using a statistical model constructively is pulling out a few numbers that correctly highlight the most important features of the table for the decision at hand.

2.2 Aside: continuous random variables

As you probably know, variables following the 'normal distribution' are one famous type of continuous random variable. The possible outcomes in this case are real numbers from minus infinity to plus infinity, $(-\infty, \infty)$. The probability of any range of outcomes is characterized by the famous bell curve (Fig. 1).

Fig. 1: Normal distribution (mean 0, variance 1)



Where the bell curve is higher it means that the associated values on the vertical axis occur with higher probability. For example, in Fig. 1, values in a neighborhood of 0.9 are about twice as likely to be observed as those in a neighborhood of -1.5. We know this because the bell curve is twice as high above 0.9 (line B) as it is above -1.5 (line A).

While it is important to understand these basics regarding continuous random variables, everything we will do in this class is most simply discussed using discrete random variables, so we will stick with the discrete case.

3 Summarizing likely outcomes of random variables

Let's return to the simpler case of modelling the outcome of the roll of a fair 6-sided die:

Table 1 (again): Prob. distribution for fair die

pr	outcome
1/6	1
1/6	2
1/6	3
1/6	4
1/6	5
1/6	6

You should be familiar with many standard concepts for summarizing a distribution: the mean, median, mode, etc. These are all measures of 'central tendency.'

Mode: the outcome or outcomes with the highest probability. This may not be uniquely defined as, say, with a fair die—all outcomes are equally likely.

Median: a value such that the probability of a lower outcome and the probability of a higher outcome are each 0.5.

Mean, expectation: The value that would occur on average from a large number of realizations of the random variable.

Detail on the mean. For a discrete random variable, x with possible realizations $r_1, \ldots, r_N$ occurring with pr. $\text{pr}_1, \ldots, \text{pr}_N$, the mean is

$$\begin{aligned} Ex &= r_1 \times \text{pr}_1 + \ldots + r_N \times \text{pr}_N \\ &= \sum_{j=1}^{N} r_j \times \text{pr}_j \end{aligned}$$

The mean is also called 'the expectation' and the expression 'Ex' stands for 'the expectation of the random variable x.' We also sometimes write x^e for the same thing.

Table 2: Computing the mean for fair die

pr	r	pr $\times r$
1/6	1	1/6
1/6	2	2/6
1/6	3	3/6
1/6	4	4/6
1/6	5	5/6
1/6	6	6/6
sum		3.5

The mean is the average value that would be observed over a large number of realizations

3.1 A comment on the mean and the term 'expectation'

The mean of a random variable is also called the expectation. Of course, we also have an everyday language meaning of expectation. For example, I expect you to do the reading for the course. A couple of standard confusions arise from the the two possible uses of expectation.

For a fair 6-sided die, the expectation is 3.5. Of course, this is not the 'expected value' in the everyday sense: we'll never see 3.5 come up. This outcome is simply the average value we would expect to obtain over many draws.

We will also talk about what agents 'expect' for financial returns. It is a hotly debated topic how closely 'expectations' that people have in practice conform to the statistical notion of expectation.

In this class, I'll try to be clear about what sense I am using. When we are doing formal analysis with equations, you can be pretty sure we are talking about the formal mathematical-statistical notion of expectation.

3.2 Expected return on a defaultable zero-coupon bond

Suppose the market price is $P = \$0.94$ for a bond that returns \$1 in 1 year. The yield to maturity is the interest rate that equates the present value of the bond's stream with the current price:

$$P = 1/(1+y)$$

or

$$y = \frac{1}{0.94} - 1 = 0.064$$

or 6.4 percent. Suppose that the bond issuer will default with probability 0.05. In default, the bond holder only recovers \$0.20 (gets 20 cents on the dollar). Thus, the payoff is a random variable with two outcomes, 1 and 0.2 which happen with probability 0.95 and 0.05, respectively. What is the expected return?

All our notions of return come down to some version of

$$1 + \text{return} = \frac{\text{future proceeds}}{\text{cost today}}$$

We define expected return in a natural way as:

$$1 + \text{expected return} = \frac{\text{expected proceeds}}{\text{cost today}}$$

For the bond then,

$$1 + i^e = \frac{0.95 \times 1 + 0.05 \times 0.2}{0.94}$$

or 2.1 percent.

The expected return is considerably below the yield to maturity. Intuitively, the yield to maturity is premised on equating the current price to the present value of *the full contractual stream of payments.* When we don't expect to receive the full contractual stream, the expected return will be lower than the yield to maturity.

3.3 Measures of dispersion

Since we don't 'expect' to obtain exactly the 'expected value,' we need some characterization of how likely we are see outcomes far from the mean. That is, we care not only about the central tendency of returns but about the dispersion of possible returns. How likely is it that I lose all my money? How likely is it that I earn twice the 'average' return?

As with central tendency, there are several standard measures of dispersion. We begin with the variance.

The variance of a random variable is defined as

$$\begin{aligned} \text{var}(x) &= d_1^2 \times \text{pr}_1 + \ldots + d_N^2 \times \text{pr}_N \\ &= \sum_{j=1}^{N} d_j^2 \times \text{pr}_j \end{aligned}$$

where $d_j = r_j - Ex$. The new variable, d_j, tells us how far the j^{th} realization *deviates* from the mean. We square these ds making them all are positive. After squaring the ds, we weight them by the probabilities and sum. Thus, variance is the mean of the squared deviations.

The variance of the outcome from the fair die is 2.92 (Table 3).

Table 3: Variance for fair die

pr	r	pr $\times r$	d	d^2	pr $\times d^2$
1/6	1	1/6	-2.5	6.25	6.25/6
1/6	2	2/6	-1.5	2.25	2.25/6
1/6	3	3/6	-0.5	0.25	0.25/6
1/6	4	4/6	-0.5	0.25	0.25/6
1/6	5	5/6	1.5	2.25	2.25/6
1/6	6	6/6	2.5	6.25	6.25/6
sum		3.5			2.92

If the units on the random variable are dollars, for example, then the units of the variance are dollars-squared. It is often convenient to consider the square root of the variance, which brings the units back to dollars. This measure is known at the standard deviation:

$$\text{std}(x) = \sqrt{\text{var}(x)}$$

3.4 Other measures of dispersion

As with central tendency, there are many measures of dispersion around the mean. One alternative dispersion measure is the *interquartile range* which is the distance between the 75^{th} and 25^{th} percentile values.

Remember, the k^{th} percentile is the value, v, such that k percent of the outcomes are smaller than v.[3]

The median, then, is the 50^{th} percentile value.

If we have a large number of outcomes, about 50 percent of the outcomes should lie between the 25^{th} and 75^{th} percentile outcomes. Thus, the interquartile range gives us a sense of how wide a range we have to take in order to be likely to capture half of the likely outcomes.

3.5 Value at risk

Returns that are far below the mean may be of particular interest to people—if returns are too low, you go bankrupt or go hungry, etc. Thus, some measures focus particularly on deviations below the mean. One measure used heavily in practice is 'value at risk.'

Suppose our random variable, x, measures the dollar loss or gain on our portfolio. The 5 percent value at risk is the value, v, such that my losses will

[3] If your test score is at the 89^{th} percentile, 89 percent of the scores are lower. *Note: I expect every one of you to score at least at the 89^{th} percentile in this course.* (That's a joke.)

exceed v only 5 percent of the time. Put another way, with high probability (95 percent) I won't lose more than v. In this context, the 5 percent value at risk is merely the 5^{th} percentile of the random variable measuring losses.

In practice, we need to specify how long a period of time we are covering. That is, I will only lose more then v *in the next day* 5 percent of the time. Thus, value at risk comes with a time horizon and a probability.

4 Joint distribution and diversification

As already emphasized, in finance we often are concerned with the outcomes for portfolios of assets. That is, we are interested in the simultaneous joint outcome for the many random variables comprising the returns on all the individual assets. As you know, a key idea is 'portfolio diversification,' whereby one is better off spreading ones invested wealth across many assets. We are building toward the mathematical underpinnings of the advice, 'don't put all your eggs in one basket.'

Let's start with some intuition. Suppose I pay a dollar to play the following game: we toss a fair coin and I get $3 dollars for a heads and zero for tails.

My expected winning is $1.50. The variance of my winnings is

$$2.25 = (-1.50)^2 \times 0.5 + (1.50)^2 \times 0.5$$

.

Now suppose instead I go halves with my friend on 2 plays of the game. We each contribute 50 cents toward each play and each get half the proceeds. This still costs me a dollar and my expected winning is still $1.50 (my expected winning from each play is now 75 cents).

The variance? There are now 4 possible outcomes for the two flips: HH, HT, TH, TT. Each outcome happens with probability 1/4 and the outcomes pay me 3, 1.5, 1.5, 0, respectively.

Apply the variance formula (for each outcome subtract the mean, multiply by the probability, and sum):

$$1.125 = (1.5)^2 \times 0.25 + (-1.5)^2 \times 0.25$$

I have cut the variance in half by 'diversifying' across 2 plays instead of going with a single play.

Instead, I might try splitting 3 ways on 3 plays, and so forth. My mean return stays the same, but my variance falls. By listing all the outcomes

and their payoffs and probabilities, you can verify that the variance splitting across N plays is,

$$2.25/N$$

If I split across a very large number of flips, I have essentially no variance whatsoever. For a large number of flips, almost exactly 50 percent will be heads and 50 percent tails. Thus, my share of the winnings will be almost exactly \$1.50. Note: spend \$1 for a 50-50 chance of zero and \$3 sounds like a decent deal, but spending \$1 to get \$1.50 for certain is a very good deal indeed.

The key assumption we have made (implicitly) here is that the coin flips are statistically independent. That is, the outcome of one flip has nothing to do with the outcome of the others. When returns are independent, the power of diversification for reducing risk accumulates very quickly. That is, the variance falls quickly as we split our investment across a broader range of gambles.

The case when the returns are not fully independent—when the asset returns move up and down together—is more subtle. We now move to this case.

4.1 Joint distribution

A multi-variate random variable is just a collection of random variables that may be related. Our 50 stocks discussed above provide an example. To define the distribution of a multi-variate random variable, we simply list all possible realizations and their probabilities. But now each realization gives the value of each random variable in the collection.

For example, if (x, y) stand for a pair of fair dice we have 36 equally likely realizations:

$$\begin{gathered}(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\\(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)\\(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\\(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\\(5,1),(5,2),(5,3),(5,4),(5,5),(5,6)\\(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\end{gathered}$$

In each pair, (x, y), the first number gives the realization for the first die, and the second give the realization for the second. Each outcome has probability 1/36.

More generally, a discrete m-variate random variable can be described in a table where each row gives one outcome for all m variables and the probability that the outcome in that row will be observed. The following describes the joint distribution of 3 random variables, x,y,and z:

Table 4: The joint distribution of 3 random variables

pr	r_x	r_y	r_z
1/3	5	-2	12
1/6	0	1	0
1/3	17	1	0
1/6	0	-3	0
mean	7.33	-0.66	4
var	50.9	2.9	32

The interpretation is that, for example, with probability 1/3 we will see the realization $x = 5$, $y = -2$ and $z = 12$ (row 1). The bottom rows give the mean and variance of each random variable.

Suppose the columns of the table labelled $r_\cdot$ give the distribution of possible values in 1 year's time of \$1 put in the asset in question, x, y, or z. Instead, if I put 50 cents in the asset instead, my value in a year will be half as much, and so forth.

I invest \$1 total, putting half in x and 1/4 each in y and z. What will be the mean and variance of my portfolio in a year's time.

To formally analyze this portfolio, we define a new random variable, q, giving my portfolio value; the new variable takes on realizations, r_q.

By definition, the realization of the value in a year's time is:

$$r_q = 0.5r_x + 0.25r_y + 0.25r_z \tag{1}$$

where r_x is the outcome for x and similar for r_y and r_z.

I can compute the mean outcome by making a table of the possible realizations for q and then applying the formula for the mean:

Table 5: Mean of the portfolio

pr	r_x	r_y	r_z	r_q	$r_q \times$ pr
1/3	5	-2	12	5	1.66
1/6	0	1	0	0.25	0.04
1/3	17	1	0	8.75	2.91
1/6	0	-3	0	-0.75	-0.13
mean	7.33	-0.66	4	**4.5**	
var	50.9	2.9	32		
sum					**4.5**

You should be able to verify that mean is given by

$$q^e = 0.5x^e + 0.25y^e + 0.25z^e$$

That is, in order to get the portfolio mean, I can just apply the portfolio weights to the means of the 3 random variables.

For any set of random variables $x_1, \ldots, x_n$, and portfolio weights, $\lambda_1, \ldots, \lambda_n$,

$$E \sum_{j=1}^{n} \lambda_j \times x_j = \sum_{j=1}^{n} \lambda_j \times x_j^e$$

In words, the mean of a weighted sum of random variables is the weighted sum of the means of the random variables.

4.2 Variance of a portfolio and diversification

The magic of portfolio diversification is essentially that while the mean of the weighted sum is the weighed sum of the means, the variance of the weighted sum is generally much less than the weighted sum of the variances.

The variance of the portfolio is given by the applying the variance formula to the realizations for q in Table 5. Thus, compute the deviations from the mean for q, square them, weight them by the probabilities and sum:

Table 6: Variance of the portfolio

pr	r_x	r_y	r_z	r_q	$r_q - q^e$	pr $\times (r_q - q^e)^2$
1/3	5	-2	12	5	0.5	0.083
1/6	0	1	0	0.25	-4.25	3.01
1/3	17	1	0	8.75	4.25	6.02
1/6	0	-3	0	-0.75	-5.25	4.59
mean	7.33	-0.66	4	4.5		
var	50.9	2.9	32	**13.7**		
sum						**13.7**

Notice that the variance is far lower than would be predicted by simply taking a weighted average of the variances of the underlying assets ($0.5 \times 50.9 + 0.25 \times 2.9 + 0.25 \times 32$). This follows *even though the random variables in Table 6 are not independent.*

But wait. How do I know that, say, x and y are not independent?

The realizations of independent random variables are totally unrelated. Thus, if I tell you the value of the outcome for say, x, this does not help you at all in figuring out the likely outcome for y.[4] If the random variables, instead, are dependent, then if I tell you something about the realization for one of

[4]With 2 independent flips of a fair coin, if I tell you the first flip came up heads, this tells you nothing about the likely outcome of the second flip.

the random variables, this helps you narrow down the likely realizations of another.

In our example, you can see that the random variables are dependent. For example, if I tell you that $r_y < 0$, you know the realization must be from the first or fourth row. Thus, you can deduce, e.g., $r_x < 17$.[5]

When the outcomes for two random variables are linked in this way, we need some summary measure of just how related the outcome are.

4.3 Covariance and correlation

The most standard measures of relation between two random variables are related to the notion of variance and are called covariance and correlation.

The formula for the covariance of x and y is,

$$\text{cov}(x, y) = \sum_{j=1}^{N} d_{xj} \times d_{yj} \times \text{pr}_j$$

where $d_{xj} = r_{xj} - x^e$ is deviation from the mean for the j^{th} outcome for random variable x; d_{yj} is analogously defined.

In steps, to compute the covariance of variables x and y from Table 6 we do the following. First we compute the deviations from the mean for each outcome of each variable. Then we take the product of these, $d_x \times d_y$. Then we take the mean of the result—that is we weight the products by the probabilities, and sum. Thus, the covariance of x and y in this case is 7.22.

Table 7: Covariance of x and y

pr	r_x	r_y	d_x	d_y	$d_x \times d_y$	pr $\times (d_x \times d_y)$
1/3	5	-2	-2.33	-1.33	3.11	1.04
1/6	0	1	-7.33	1.66	-12.22	-2.04
1/3	17	1	9.66	1.66	16.11	5.37
1/6	0	-3	-7.33	-2.33	17.11	2.85
mean	7.33	-0.66				
var	50.9	2.9				
sum						7.22

Notice that if both variables are above the mean (both ds are positive), then the product of d_{xj} and d_{yj} is positive. If both variables are below the

[5]For those with probability training, what we are describing in intuitive terms is that if x and y are independent, then the conditional distribution of x given y equals the unconditional. With dependent random variables, the conditional and unconditional distributions differ.

mean the product is also positive. Thus, if the variables tend to move above and below the mean together (in the same realizations; on the same rows), $d_x \times d_y$ will generally be positive; the weighted average will be positive, and, hence, the covariance will be positive. In words, variables that move up and down together have positive covariance.

If when one variable is above the mean, the other is *equally likely* to be above or below the mean, then the terms $d_x \times d_y$ will be evenly split between positive and negative and the covariance will tend to be near zero. This will be true, for example, if the random variables are independent.

If when one variable is above the mean, the other tends to be below, the $d_x \times d_y$ terms will generally be negative, and covariance will be negative.

If the outcomes tend to be high and or low together, the variables have positive covariance. If the outcomes tend to move in opposite directions, the variables have negative covariance.

4.4 Correlation

Correlation of x and y is defined as

$$\text{cor}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\,\text{var}(y)}}$$

That is, the correlation between x and y is the covariance of x and y divided by the standard deviation of x times the standard deviation of y.

Correlation is always between minus 1 and 1. (An important mathematical relation shows that

$$|\,\text{cov}(x, y)| \leq \sqrt{\text{var}\,x\,\text{var}\,y}$$

where || means absolute value.[6])

Correlation of 1—perfect positive correlation—means that the deviations from the mean of the two variables are always exactly proportional:

$$d_x = k d_y$$

for some constant $k > 0$. With perfect negative correlation, $k < 0$, so that the variables move proportionally, but in opposite directions.

[6]This relation is known as the Cauchy-Schwarz inequality and is related to the triangle inequality which states the simple fact that no 1 side of a triangle can be longer than the sum of the lengths of the other two sides.

4.5 Aside: Independence and covariance

Independence requires that the random variables are unrelated in essentially every possible sense. Thus, independent random variables must have no covariance.

Covariance measures one sense of how random variables are related. Two random variables that have zero covariance may be related in other ways. Thus, variables with zero covariance need not be independent.

That is: Independence implies zero covariance, but not vice versa.[7]

4.6 Variance of a portfolio of two assets

Above, we showed that if we have returns x and y and put portfolio shares a and b into them, respectively, then the mean return on the portfolio will be,

$$E[ax + by] = ax^e + by^e$$

The analogous relation for variance is more complicated and gives a key characterization of diversification:

$$\text{var}(ax + by) = a^2 \,\text{var}(x) + b^2 \,\text{var}(y) + 2ab \,\text{cov}(x, y) \tag{2}$$

From this formula, we can see some lessons about diversification.

Suppose i) that the variance of x and y are each V, ii) the covariance is zero, and iii) we put equal shares in the two assets so that $a = b = 1/2$.

Now the variance of the portfolio will be

$$\text{var}(0.5x_1 + 0.5x_2) = 0.25V + 0.25V = 0.5V$$

That is, the variance will be half the original. This is the result we found when moving from one coin flip to two.

Suppose instead that the two random variables had perfect (positive) correlation. In this case,

$$\begin{aligned} \text{cor}(x, y) &= \frac{\text{cov}(x, y)}{\sqrt{V \times V}} \\ 1 &= \frac{\text{cov}(x, y)}{\sqrt{V \times V}} \\ &= \text{cov}(x, y)/V \end{aligned}$$

[7] Those of you who have had statistics may remember that in the *special case* of jointly normal random variables, zero covariance does imply independence.

or $\text{cov}(x, y) = V$.

Plugging this in the formula for variance of the portfolio, (2), gives

$$\text{var}(0.5x_1 + 0.5x_2) = 0.25V + 0.25V + 2(0.25)V = V$$

When the assets are perfectly correlated, there is no benefit to diversification.

Suppose instead that the two random variables had perfect (negative) correlation. In this case, $\text{cov}(x, y) = -V$, and you should be able to verify that the equal-weighted portfolio has variance of zero.

As you are probably gathering, covariance and correlation shed important light on portfolio risk and the benefits of diversification.

4.7 Portfolios with multiple assets

We have given the formula for the variance of a portfolio of two correlated assets. Of course, generally we want more than two assets in the portfolio. The formula generalizes in a natural way to the multiple asset case, but things get a bit messier than we plan to get into in this class.

Fortunately, we can capture the key issues by thinking only of two assets.

In particular, we will think of one 'asset' as the overall portfolio. The second asset is an asset we are considering adding to the portfolio.

We will approach optimal portfolio allocation in the following way. Suppose I have a portfolio and the overall portfolio return is characterized by the random variable q. The portfolio has mean Eq and variance $\text{var}(q)$.

I consider shifting a bit of my funds out of the portfolio and into a new asset, x, with mean Ex, variance $\text{var}(x)$ and its covariance with the portfolio is $\text{cov}(q, x)$.

Our portfolio theory will address how we make decisions like this optimally.

4.8 A note on covariance vs. variance as a measure of risk

We often equate risk with variance. This is probably alright, but when we do so it is vital that we remember that there is good, bad, and indifferent risk. That is, not all variance in an individual asset is bad variance.

Suppose I have an asset that pays me $\$500,000$ in very rare events (say, with probability 0.0000001) and pays zero otherwise. This asset sounds something like a lottery ticket: it's expected return is very small and the variance of this return is very large. We might naively conclude that this asset is a poor investment.

But any decision about whether an asset is a poor or great investment *must* involve considering *covariance* of the payoff with my other assets. Concretely, if the asset just described pays off if and only if my $\$500,000$ house burns down, then it is a risk reducer for me.

We have a name for this sort of lottery ticket: homeowner's insurance.

In short, if we just look at the mean and variance of a particular asset's return, we cannot tell the difference between lottery tickets and homeowners insurance: viewed in isolation, each pays off in a very big way but only very rarely. It is covariance with stuff I care about that differentiates the two. The lottery ticket has a payoff that is unrelated to my other assets, while the insurance has a return with a large negative correlation with the value of my house.

If you want to think about risk in a particular asset, you should think about the way the variance of that asset return would contribute to variance of an overall portfolio. This, it turns out, this is largely a matter of covariance.

4.9 Which summary measure of how random variables are related?

We have many measures of central tendency (mean, median, ...) and of dispersion (variance, interquartile range, ...). Which should we use?

Of course, the answer must be, 'it depends,' or stat. books would be much shorter and I'd be in a different line of work.

One key issue in choosing among measures is sensitivity to outliers. Outliers are small probability realizations that are far in value from the vast bulk of outcomes.

Take our fair die. Change the the value on the side with 6 dots to 60 instead. With probability $5/6^{th}$s, the outcome is between 1 and 5. With probability 1/6, the outcome is about 10 times this larger.

Note that the mean is now 12.5 instead of 3.5, but the median has not changed at all since the 50^{th} percentile is unmoved. The mean is fairly sensitive to outliers, whereas the median is quite insensitive.

The variance is now 452.92 instead of 2.91, but the interquartile range has not changed at all. The variance is extremely sensitive to outliers because we square the deviations. (Squaring magnifies larger values much more than smaller ones: ($1^2 = 1$, $10^2 = 100$, $100^2 = 10,000$).)

The interquartile range and median are not sensitive at all to outliers in our example, but the mean is sensitive and the variance is very sensitive.

So is sensitivity to outliers good or bad? It depends. In some contexts, we want to ignore outliers, in others they are the main point of the exercise.

Suppose the random variable describes snowfall in Baltimore. We have a client deciding whether or not to buy a Volvo L45F front loader (Fig. 2) to shovel her driveway.

Figure 2: Volvo L45F

You are a statistical consultant deciding what summary measures of the distribution of snowfall in Baltimore would be most helpful to the client in deciding whether to buy the L45F. This will depend importantly on the objectives of the client.

The client tells you that the decision should driven mainly by the issue whether the L45F would be needed often in a typical year. In this case, a measure of central tendency and dispersion of snowfall that is insensitive to outliers is probably in order. Most years, the number of times one would need the L45F to clear ones driveway in Baltimore is probably near zero.

On the other hand, suppose the client is running an emergency care facility that must remain open at almost any cost. In this case, a summary measure that emphasizes snowfall outliers may be exactly what is in order. You want to be sure to be able to clear the driveway even in the worst cases.

This discussion raises a few important points. Often analysts are called on to summarize random variables for some policymaker or manager. One important family of issues regards which measure will *create the proper impression* Which measure might be misleading? Which measure best high-

lights the important features? The answer to these questions is very dependent on the context and the particular question being considered.

All these issues are highly relevant to the question of risk management in financial institutions. In some cases, we may be most interested in what our returns will be on a typical day (ignoring freak events like blizzards or a meltdown of the financial system); in others, the low probability, freak, events may be the purpose of the exercise.

5 And finally, Emerson, Knight and garbage in garbage out

Formal tools are immensely valuable. They provide a rigorous framework in which to think consistently about complex issues. As Ralph Waldo Emerson noted, however, 'a foolish consistency is the hobgoblin of little minds adored by little statesmen and philosophers and divines.' In the statistical arena, a related, but less elegantly stated, thought is 'garbage in, garbage out.'

No matter how sophisticated your statistical model, if you did not get the list of possible events and their likelihoods about right, you will get bad (but internally consistent) guidance out.

In the 1920s, the great economist Frank Knight emphasized the difference between uncertainty where the outcomes and probabilities were all known versus more profound uncertainty when they are not. He called the case where the outcomes and probabilities are known, 'risk', whereas the deeper sense cluelessness was 'uncertainty.' This is a very important distinction.

(Let me emphasize that while Knight's distinction is important, his use of these terms has not been widely adopted; thus, 'risk' and 'uncertainty' are often used interchangeably. For clarity, folks referring to what Knight called 'uncertainty' we will often use the term 'Knightian uncertainty.')

The formal statistical tools of finance go awry when we pretend that we know all the relevant outcomes and probabilities when, in fact, we face Knightian uncertainty. That is, we apply the formal statistical tools just described, while overlooking the fact that we do not really have a reliable list of all outcomes and the associated probabilities.

In many risk models before the financial crisis, the probability of a large nationwide decline in house prices was zero or nearly zero.

There is another version of this problem that many folks think is often at work in financial crises. In particular, financial firms form estimates of the covariance of various returns based on data from how the economy performs in normal times. Based on these estimates, the firms choose an 'optimally

diversified' portfolio. Based on the historical data (most of which comes from normal times), the firm finds that assets returns have low covariance, so that the effects of diversification will be powerful. That is, spreading funds over a large number of stocks should eliminate most variance from the portfolio return.

Under certain stressful conditions (conditions we *ex post* label a crisis), all asset returns begin moving together. Correlation of all risky returns rises toward 1, the benefits of diversification evaporate, and all of the sudden, the 'optimal portfolio' is very risky.

The statistical modellers in this case will report that this outcome 'is nearly impossible,' 'we couldn't have known.' It was a freak event, like seeing a black swan. A sudden sharp rise in correlation is one example of a 'black swan phenomenon' of the type popularized by Nicholas Taleb.[8] Some think that such changes are often at the heart of financial crises.

A more realistic perspective (and one less likely to blow up the economy) is that the mechanism driving the economy and financial markets is quite complicated and events like a sharp rise in correlation may be quite likely under certain conditions. Since we do not observed these conditions very often, our assessment of their probabilities may be quite unreliable.

One lesson is that when you see a black swan, you can either imagine that you are experiencing a very rare event, or that you have entered a region of the world in which black swans are common. The latter possibility is important, because we might be able to learn the signs of such regions and thereby prepare for the event.

Overall, our conventional formal statistical tools require an exhaustive list of outcomes and their associated probabilities.[9] In practice, we sometimes (often?) face unknowns where the full set of outcomes is unclear and/or we don't have a very good idea of the relevant probabilities. That is, we cannot quantify the way values will be resolved in the manner required for us to reliably use our formal statistical tools.

Let me end where I began:

Managers, corporate boards, and regulators rely heavily on summary measures regarding the likely pattern of portfolio returns in making risk management decisions. The question of which summary statistics are most revealing of the relevant information is one of great practical importance. As consumers of statistical information, managers, boards and regulators need to know when they might be being mislead. As producers of this

[8] e.g., `http://www.nytimes.com/2007/04/22/books/chapters/0422-1st-tale.html`.

[9] There are various less conventional means for dealing with less complete knowledge.

information, individuals supplying this information need to know how most clearly to summarize the situation as they see it.

Many folks see breakdowns at every level here playing a role in the recent crisis. One goal of these notes is to help you begin building some perspective on the practical uses and abuses of statistical concepts. I hope you all will leave here less likely to blow up your company or the economy than earlier generations.