Basic Probability and statistics for Finance
266: Financial Markets and Institutions

Jon Faust

http://e105.org/e266

March 28, 2017

▶ **Risk so far in this class...**

- So far we have only flirted with one of the thorniest and most important issues in finance: how do markets deal with risk?
- We've mainly appealed to the rule of thumb that potential holders of risky assets demand a yield premium in order to hold risker assets.

▶ **...** The portfolio problem for any individual, group or firm is this: How should I optimally divide my investable wealth across the myriad assets in the market?

▶ **Eggs and baskets and diversification**

- One useful rule of thumb is this: don't put all your eggs in one basket.
- The essence of this folk wisdom is that diversifying your portfolio will be less risky
- But...

▶ **Diversify how?**

- The folk wisdom raises as many questions as it answers:
  - How many baskets should I use?
  - And since eggs have different sizes and colors, which eggs in which baskets?
- Portfolio theory provides a framework for answering the investment equivalent of these questions

  Which assets should I buy? And how much of each asset?

▶ **Models, math and complicated problems**

- There are of course zillions of assets out there in the world
- Giving sound advice requires rigorous treatement of risk
- Thus, we apply some tools from probability and statistics.

▶ **This course**

- In this course, I will introduce these tools and present some of the simplest results

  real problems require you to go further: the corporate finance and investments courses are a natural next step.

- I'll also provide a framework for thinking about the full rigorous approaches used on Wall Street.

  This perspective applies both to the simple tools I'll present and to the most complex approaches.

▶ **The essence of risk modelling**

- Formal risk modelling begins with writing down a list of all possible outcomes for the risky event in question
- And assigning a probability to each event

  This probability is the likelihood that the outcome in question will be the one realized.

▶ **Example 1: Rolling a 6-sided die**

Prob. distribution for fair die

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

▶ **Risk models generally**

- At the heart of any risk model is a table like that just given
- Each row represents a possible outcome and is assigned a probability.

▶ **A risk model is nothing more or less than a table listing all relevant outcomes along with their probabilities.**

▶ **A first risk model in finance**

- Suppose I have a portfolio stocks in the S&P 500: IBM, Merk, . . . , Apple.
- Suppose my risk model characterizes the price of these shares at the close of business tomorrow.

▶ **My 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

▶ **Issues**

- This table has 501 columns

  500 stock prices in a given outcome and the probability of this outcome

- This table has zillions of rows.

  There are myriad combinations prices these stocks could take on tomorrow.

- We're going to need some tools to condense this table into a usable form.

▶ **Issues**

- We'd like to find ways to summarize this table that highlight the most important issues from the standpoint of choosing a portfolio.
- And that leave out stuff that is of little or no importance.
- Portfolio theory helps us pick out summary measures that are likely to be most useful.
- But as we'll return to at the end, just what is the 'right' summary is open to question

▶ **With that introduction, let's learn some formulas.**

▶ **Probability and Statistics tools**

- I'll start with some probability and statistics tools that you should all know already.

▶ **Summarizing typical outcomes and dispersion**

- Focus for a moment on the outcome for a single item: the role of a die, flip of a coin, or price of a single stock.
- The first two natural features to summarize are the most likely outcomes and how dispersed the possible outcomes are around these most likely outcome.

▶ **Standard summary measures**

- Central tendency: what values are most typically observed
- Dispersion: What kind of spread in outcomes do we see

▶ **Central tendency**

- Name some conventional measures of central tendency

  Mean, median, mode

▶ **Definition: mean**

- Take an r.v., x, with realizations $r_1, \ldots, r_n$ with associated probabilities $\text{pr}_1, \ldots, \text{pr}_n$.
- The mean of x is defined as
$$x^e = \sum_{j=1}^{n} \text{pr}_j \times r_j$$
- A weighted average of possibilities where the weights are probabilities

▶ **Notation/terminology**

- The mean of x is also known as the expectation of the random variable is often written as $E[x]$ and x^e.

▶ **Example**

- The mean of the r.v. describing the roll of a fair die is

  3.5

- Note, clearly the mean is not the 'expected valule' in the simple everyday sense: we never expect to roll a 3.5.

▶ **Median**

- The median is a value such that the probability of a lower realization is 50 percent (and similarly for a higher realization)
- For the die, any value $3 < m < 4$ satisfies this.

  We use some convention such as picking the mid-point of the range to call the median

▶ **Average in greater depth.**

- If I have a set of numbers, $\{7, 23.5, 3.14, 52\}$.
- The average is the sum divided by the number of elements:

$85.64/4 = 21.41$

▶ Average vs. weighted average

- Average formula:
$$\sum_{j=1}^{N}(1/N) \times x_j$$
- In an average, each element gets the weight $(1/N)$.
- In a weighted average we replace $(1/N)$ with a potentially different weight, w_j, for each of the xs:
$$\text{weighted ave.} = \sum_{j=1}^{N} w_j \times x_j$$
where the w_js sum to 1.
- So long as the weights themselves add up to 1, we call this a 'weighted average.'
- Many of our formulas in finance are about weighted averages
- Duration: the weights are present value shares
- mean and variance (today)

  the weights are probabilities
- portfolio expeced returns

  the weights are the share of funds placed in a given asset
- Thus, we will do a lot of summing up a list of values where those values will be weighted by something.

▶ Summary measures of dispersion

▶ Dispersion

- As with central tendency, there are many dispersion measures

  they generally summarize how realizations tend to deviate from the mean
- Most important measure here is 'variance'

▶ The variance

- The outcomes for x are $r_1, \ldots, r_J$.
- For each possible realization or outcome, define the 'deviation' of that realization from the mean:
$$d_j = r_j - x^e$$

- If you were expecting x^e and got outcome r_j instead, then d_j would be a measure of how much better or worse the outcome is from the expected
- Variance of x

$$\text{var}(x) = \sum \text{pr}_j \times d_j^2$$

- Thus, the variance is a weighted average of the squared deviations.

were the weights are the probabilities of the deviations.

▶ **Variance for a 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

▶ **Definition: standard deviation**

- The standard deviation of a random variable is the square root of the variance.

$$\text{std}(x) = \text{var}(x)^{1/2}$$

▶ **Aside:: Units, standard deviation**

- In computing the variance, we square the values of the original variable.
- If the units of x were, say, dollars, the units of the variance are dollars-squared.
- The standard deviation returns things to the units of the original variable and is, thus, sometimes easier to interpret than the variance

▶ **Another measure of dispersion: interquartile range**

- Another standard measure of dispersion is the interquartile range

$$\text{interquartile range} = 75^{th} \text{ percentile } - 25^{th} \text{ percentile}$$

▶ **Aside:: percentile**

- The k^{th} percentile is a value such that k percent of the outcomes are below this value

- Thus, the median is also the 50^{th} percentile

▶ **Interquartile range**

- By definition, over many realizations, 50 percent of outcomes should fall within the interquartile range
- Thus, you have a 50-50 chance of falling outside the interquartile range

▶ **Finance**

- We have spoken of expected returns and expected inflation
- Now we have a formal treatment of what this means that can be used in formal models.
- We have talked of risky asset returns.
- Some aspects of risk can be summarized using the various measures of dispersion.

▶ **Two motivating examples**

▶ **Example 1: A simple game**

- A game: I pay \$1 and based on a fair coin toss I get either \$3 or \$0.
- My expected payoff is

  \$1.50

- Variance of my payoff?

  The deviations from the mean are \$1.50 and −\$1.50, each with pr. 0.5. The variance is

  $$2(1.5^2 \times 0.5) = 1.5^2 = 2.25$$

▶ **Diversification**

- Suppose a friend and I go halves on two plays of the game

  we each pay \$1, and split any winnings evenly.

- Possible outcomes on the flips are

  HH, TH, HT, TT, each has pr 0.25

- My winnings for each of these realizations:

  $$3, 1.50, 1.50, 0$$

- My expected winnings?

1.50, same as with a single play

- The variance of my winnings?
- Deviations are 1.5, 0,0, 1.5

  So the variance is,

$$2(1.5^2 \times 0.25) = 1.5^2/2 = 1.125$$

- I cut my variance in half by 'diversifying' my \$1 investment equally over two plays.

▶ **More diversification**

- Suppose I get together with 2 friends and and pool the winnings on 3 plays.

  Mean stays the same, variance goes down

- Having exhausted my list of friends :(I decide to pool with strangers.

  Thus, I pool with a large number of folks, each pays for one ticket, but we split the winnings evenly.

▶ **The society-wide pool**

- Expected winnings for each of us?

  \$1.5.

- Variance goes to zero as we increase the number of plays we pool

  Intuition: for a large number of flips, almost exactly 1/2 will be heads and 1/2 tails.

▶ **Aside:: Varaiance for large number in the pool**

- The formula for the variance of the winnings when diversified over N plays of this game is,

$$2.25/N$$

- see the prob. and stats. reading

  this is a version of the law of large numbers from statistics.

▶ **Lesson: by splitting our investment over various risky assets, risk as measured by variance falls, but expected return remains a weighted average of the underlying expected returns.**

▶ **Example 2: types of lottery ticket**

- The relevant notion of risk involves both the dispersion of possible outcomes AND MORE IMPORTANTLY how the outcomes of different assets are related.

▶ **Reminder**

- A lottery ticket is an asset that almost always pays zero, but occasionally pays, say, $500,000.
- Insurance against fire damage is as well: it almost always pays nothing, but occassionally pays, say $500,000.
- Both of these may have the same expected return and variance.

but one is a foolish investment, the other is generally a wise investment.

▶ **Lesson**

- By looking at the outcomes for one asset in isolation, you can never know whether the variance in its return is good risk that agents will pay to face or bad risk that they must be enticed to face.
- Put another way: the riskiness of an asset must be assessed in terms of co-movement—that is, how its return varies with other asset returns.

▶ **Thus, we need summary measures of co-movement**

▶ **Example:**

The joint distribution of 2 random variables

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

▶ **Measures of how multiple outcomes move together**

- Covariance is related to variance and is one main measure of how variables move together
- Take two r.v.s x and y with realizations r_{xj} and r_{yj} which happen with probability pr_j.
- Define deviations from the mean for each variable: d_{xj}, d_{yj}
- Covariance:

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

▶ **Example**

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
sum						7.22

$x^e = 7.33$, $y^e = -0.66$.

▶ Intuition for covariance

- If when x is above its mean, y tends to be as well, and when x is below its mean, y tends to be as well ...
- Then $d_{xj} \times d_{yj}$ will predominantly be positive
- Thus, covariance will be positive
- If when x is above the mean y tends to be below, and vice versa,

  then $d_{xj} \times d_{yj}$ will predominantly be negative
- And so covariance will be negative

▶ Correlation

- Correlation of x and y is defined as

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

- One can show that correlation is covariance scaled to fall between -1 and 1

  (correlation must be in this range for reasons tied up with the triangle inequality)
- When corr. is 1, the assets move in lockstep, one-for-one with each other

▶ Diversification, covariance, and the examples

- What is the covariance of the payoff of the lottery ticket with other outcomes in society?

  Zero.
- What sign is the covariance of the payoff of fire insurance with the value of my house?

  Negative. Insurance payoff is high when house burns down.

▶ That's it for the primer on statistics

▶ **Some general practical discussion of covariance**

- The returns on the equity of two firms in any industry probably have (pos.,neg.,zero) correlation?

  In principle, could go either way.

- What forces might push the correlation toward the negative?

  The firms compete and the gains of one could be the losses of another.

- What forces might push the correlation toward the positive?

  All subject to the same market conditions for that industry. Newspapers all lost value with the rise of the internet, essentially all bank stocks rose after Trump's election.

▶ **The facts**

- Empirically, returns across firms in an industry are generally measured to be positive, but only slightly.
- The correlation tends to be near zero because much of the variance of the return of individual firms is due to idiosynratic stuff.

  batteries catch fire (Tesla, Samsung), Steve Job's dies (Apple), you choose to systematically fake emissions tests (VW), ...

▶ **Bottom line**

- A risk model is a big table: each row is a possible outcome for all variables of interest along with the probability that this will be the outcome realized.
- This table is large and unweildy and we are forced to summarize it in much simpler ways
- Main statistical summary techniques are about
  - **central tendency**

    mean, median, mode,...
  - **dispersion**

    variance—and its relatives std. dev. and correlation—and interquartile range,...
  - **co-movement**

    we only mentioned covariance

▶ **Bottom line**

- From a substantive standpoint: risk is very importantly about co-movement and not just dispersion.

- Ignoring co-movement, lottery tickets and house fire insurance look like equally foolish investments.

▶ **Finally a word from Warren Buffett**

- Warren Buffett famously said 'beware of geeks bearing formulas'
- In this context, how do we apply this wisdom?

▶ **Models and summaries**

- We have this incomprehensibly big table.
- We want the best way to compactly or briefly summarize it
- You can properly think of this as creating the SparkNotes version.

  Much shorter focussing on just the important stuff.

▶ **Life lesson/warning: Condensing things down to just the important bits is always perilous**

▶ **Example 1.**

- Tolstoy's War and Peace is assigned in Sally's lit class, but she doesn't have time to read it.
- Great idea! Just read the SparkNotes

▶ **Question: Sally...**

- a) will probably do O.K. on the exam.
- b) will almost certainly fail.
- c) faces a nontrivial risk of catastrophic failure.
- d) a & c
- In my experience the answer is (d).
- A *high quality* summary may be adequate for many purposes but any summary has strengths and weaknesses, so relying on a summary can be dangerous.

  And you better understand the strengths and weaknesses

▶ **Example 2: financial risk modelling**

- We build a risk model (that big table we've been discussing) to help us diverify our funds across stocks in the S&P 500.
- We summarize its contents as a basis for our portfolio decisions

- If we formed a high quality risk model

  That is, we got the possible outcomes and probabilities about right

- And if we formed a high quality summary,
- Then, our summary will provide a solid basis for portfolio choice.
- But it can be very hard to tell if the model is of high quality

  And if not, we are subject to the garbage in/garbage out critique

- And as in the SparkNotes case, we know that any given summary wil have strengths and weaknesses.
- Thus, beware of Geeks bearing formulas.
- More constructively: when relying on a risk model, be sure you understand both the reliability of the 'table' and the strengths and weaknesses of the summary measures you are looking at.

▶ **Ignore geeks bearing formulas?**

- The fact that it is hard to make a good risk model and use it constructively does not mean you should simply not bother.
- The alternative is making decisions without systematically assessing risks

▶ **As Trish Little emphasized: history has made pretty clear that some mixture of judgment, intuition, and formal modelling is the way to go.**