Statistics 101

A

Dealing with uncertainty—experienced as randomness—is fraught with limitations, yet we can get at least *some* formal handle on it.

We can, first, simply observe random outcomes repeatedly and essentially count what happens. On 1000 tosses of some coin, we might observe, e.g., 104 head and 896 tail outcomes. We may interpret the frequencies of occurrences (104/1000 and 896/1000) as indicators of the likelihood of the two outcomes.

We also conclude that this particular coin does not seem to be very fair—that's because we have an idea of how a coin ought to behave. We intuit the chance or *probability* of observing either head or tail as about 50%, in the sense of counts or frequencies we would expect to see. With that idea of a fair coin in mind, we would not have been surprised to see, say, 495 occurrences of head. Here we deduce: this coin is rigged.

These two approaches of (i) watching and counting and of (ii) thinking and inferring are referred to as statistics and probability theory, respectively. They most often work in tandem. In a poll, for example, we count the responses of a small set of people to estimate some overall opinion; by making some assumptions about the nature of the involved uncertainty, we can then try to infer the confidence we can put in our estimate.

A.1 Random Variables and Probabilities

To formalize our view on randomness, we start off with the concept of a so-called *random variable*. Think of it as an entity or device that produces one specific output: a single number. It simply selects one number out of many, by chance. How likely a given number is bound to occur is governed by probabilities.

The random variable X describing a die, for example, can result in one of six numbers $1, 2, 3, 4, 5$, or 6, each with a probability of $\frac{1}{6}$. This is an example of a *discrete* random variable.

M. Auer, *Hands-On Value-at-Risk and Expected Shortfall*, Management for Professionals, https://doi.org/10.1007/978-3-319-72320-4

We can also inscribe larger numbers on our die, e.g., 10, 20, 30, 40, 50, and 60. This pimped die will help us think of how to label stuff. The first outcome, i.e., 10, is called x_1, and its probability of $\frac{1}{6}$ is variously called p_1 or $p_{X=x_1}$. Overall, x_i is the outcome $10 \times i$, with $p_i = \frac{1}{6}$. (The standard die, where x_i corresponds to the outcome i, unhelpfully blurs name/index and number/outcome.)

First and obviously, the probabilities involved must always sum up to 1:

$$p_1 + \cdots + p_6 = \sum_{i=1}^{6} p_i = \sum p_i = 6 \times \frac{1}{6} = 1 = 100\%.$$

The probability of observing an outcome larger than 25 is

$$p_{X=x_3} + p_{X=x_4} + p_{X=x_5} + p_{X=x_6} = p_3 + p_4 + p_5 + p_6 = \sum_{i=3}^{6} p_i = \frac{4}{6} = 66.6\%.$$

When throwing the die twice in a row, the probability of observing 10 followed by 50 is

$$p_{X=x_1} \times p_{X=x_5} = p_1 p_5 = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

—unsurprisingly, as there are 36 possible combinations.

Next, consider a spinning top like the one in the movie Inception. After a spin, once it falls and stands still, its handle will point in an arbitrary, random direction—this angle X is a random variable between 0 and 2π. It has uncountably many outcomes (provided we can measure the angle arbitrarily precisely), which makes X an example of a *continuous* random variable. No outcome or angle x is more likely than any other. This immediately means that we cannot assign a positive probability to an individual outcome—if we used even the smallest such probability ϵ for this, the probabilities could never sum up to 1 (because, well, $\infty \times \epsilon > 1$).

We instead capture the involved probabilities via a function $p(x)$, with $p(x) = \frac{1}{2\pi}$ for x in $[0, 2\pi]$, and $p(x) = 0$ otherwise. Why so? Well, this makes sure that the whole area under $p(x)$, the rectangle $2\pi \times \frac{1}{2\pi}$, equals 1. We can now interpret slices or partial areas over outcome ranges as probabilities. For example, the probability of X falling between 0.12 and 0.25 is the corresponding (in this case, rectangular) area $(0.25 - 0.12)\frac{1}{2\pi}$.

The function $p(\cdot)$ is called *probability density* function. It need not be constant, just positive and covering an area of 1. Slice areas or probabilities are then generally expressed as integrals, and the probability of X falling between a and b is

$$\int_a^b p(x)\, dx.$$

We let this sink in using our example. The probability of *any* outcome occurring is, in our case,

$$\int p(x)\,dx = \int_{-\infty}^{\infty} p(x)\,dx = \int_0^{2\pi} \frac{1}{2\pi}\,dx = \frac{1}{2\pi}x|_0^{2\pi} = 1 = 100\%.$$

In a clockwise arrangement, the probability of a lower-right or south-east outcome, i.e., that X lies between east $(\pi/2)$ and south (π), is the plausible

$$\int_{\pi/2}^{\pi} \frac{1}{2\pi}\,dx = 25\%.$$

So in the continuous case, we only ever really deal with outcome *ranges*. The probability of a specific outcome to occur, e.g., $X = a$, is $\int_a^a p(x)\,dx = 0$, as mentioned before.

Both the discrete and the continuous examples were rather boring, as all the outcomes were equally likely (such random variables are called *uniform*). Let's make the next example a bit more exciting. Let Z denote the time you have to wait in line at some supermarket (a precise stop watch makes this a continuous random variable). Now, you might have observed that you usually have to wait between 1.5 and 2.5 min, but rarely less than 1 or more than 3, and never longer than 4. So, first, this is clearly not uniform. Second, unlike in the examples above, we don't know the real probabilities involved—but based on our experience, we can simply invent or postulate some probabilities and try to express them via a $p(x)$. We want $p(x)$ to be 0 for $x < 0$ (we can't wait a negative amount of time), and we set $p(x) = 0$ for any $x > 4$. We want the slice areas, i.e., probabilities, around 2 min to be larger than the areas at the edges of our 4 min range in order to match our anecdotal observations. A simple way to achieve this is to shape the function like a triangle with its peak at 2 min set to $p(2) = \frac{1}{2}$, for the whole, now triangular area must again equal $1 = \frac{1}{2} \times 4 \times p(2)$. We thus have $p(x) = \frac{1}{4}x$ for x in $[0, 2]$ and $p(x) = -\frac{1}{4}x + 1$ for x in $[2, 4]$ (see Fig. A.1).

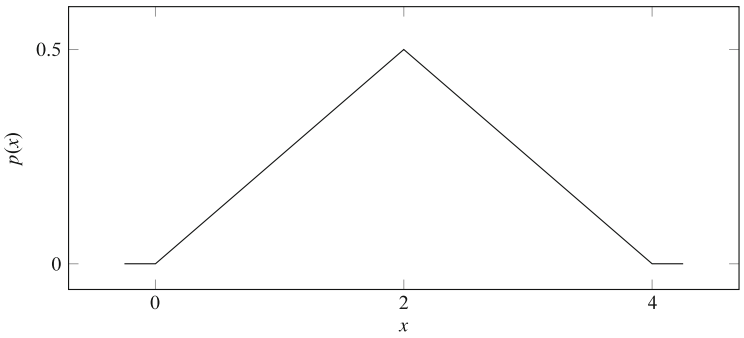


Fig. A.1 Triangular probability density

If we trust our hard-earned probabilistic model, we can now compute the probability of waiting between 17 and 25 s as the respective area under this function—the integral $\int_{17/60}^{25/60} \frac{1}{4} x \, dx$.

So outcomes and probabilities together describe and determine a random variable's behavior, its *distribution*.

A.2 Expected Value

As we have seen above, describing a random variable with all those outcomes and probabilities can be a wordy affair. We are looking for a way to get across some core characteristics of a random variable in a shorter, more succinct manner. On a hunch, we let us inspire by how we tend to average large sets of numbers (e.g., all the individual incomes of people living in Kansas) in order to compress the vast amount of information therein.

If a random variable X can have n discrete outcomes x_i, each with probability p_i, then we expect an "average," probability-weighted outcome—or *expected value*—of

$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i \, p_i.$$

If each outcome is equally likely, we have $p_i = \frac{1}{n}$, and this expression becomes the familiar average.

In case of a continuous random variable, X can take on infinitely many values; their probabilities are described via a probability density function $p(x)$. By direct analogy with the discrete case we have[1]

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \, p(x) \, dx.$$

(In both cases, or course, the probabilities themselves must always sum up to $1 = 100\%$, i.e., $\sum p_i = 1$ and $\int p(x) = 1$.)

Our examples fare as follows:

- The expected value of our pimped die is

$$\mathbb{E}[X] = \sum_{i=1}^{6} x_i \, p_i = \sum_{i=1}^{6} 10 i \, p_i = 35.$$

- The expected value of a standard die is $\sum_{i=1}^{6} i \, p_i = 3.5$.

[1] The discrete p_i corresponds to the infinitesimally small $p(x) \times dx$.

- The expected value of our spinning top is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, p(x)\, dx = \int_{0}^{2\pi} x \frac{1}{2\pi}\, dx = \cdots = \pi.$$

- The expected value of the triangle probability density of waiting times we constructed at the beginning of this chapter is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, p(x)\, dx = \int_{0}^{2} x\, (\frac{1}{4}x)\, dx + \int_{2}^{4} x\, (1 - \frac{1}{4}x)\, dx = \cdots = 2.$$

  (This corresponds to what we'd expect—the expected value here lies at our triangle's peak. As an exercise, try using a triangular probability density that is not isosceles and both guess and calculate its expected value.)

The expected value alone can't possibly give us the full picture of a random variable, yet it provides a first, brief glance at its behavior. If, for example, you hyper-pimped a die and only told me its new expected value of 3500, I might already get a fairly good impression of that die without knowing the details. (I might be wrong, because of course you could just have replaced a standard die's 6 with 20,985 to obtain that very same expected value.)

Sometimes the expected value already tells us all we need to know. Imagine a die game where you win a roll's outcome in dollars, e.g., a 4 nets you four bucks. Should you be willing to pay 3 dollars to take part in this game? We know that the die's expected value is 3.5, i.e., when playing repeatedly, you expect to receive 3.5 dollars and to thus earn 50 cents on average. Clearly, the 3 dollars investment would be worth it, but, alas, such games do not exist. If we reverse the setting, though, we obtain a game that does: would you be willing to *offer* the die game if someone paid 4 dollars to take part? Sure you would, and so do others; such games go by the name of lottery.

Moving on, the expected values intuitively extend to functions of randoms $f(X)$, i.e., to what we expect $f(x)$ to be on (probability-weighted) average, in both the discrete and the continuous case:

$$\mathbb{E}[f(X)] = \sum_{i=1}^{n} f(x_i)\, p_i,$$

$$\mathbb{E}[f(X)] = \int_{-\infty}^{\infty} f(x)\, p(x)\, dx.$$

For any constant c, we have

$$\mathbb{E}[cX] = c\, \mathbb{E}[X],$$

or, as special case (think: $X = x_1 = c$ with $p_1 = 100\%$),

$$\mathbb{E}[c] = c,$$

or, more generally,

$$\mathbb{E}[c_1 f_1(X) + c_2 f_2(X)] = c_1\,\mathbb{E}[f_1(X)] + c_2\,\mathbb{E}[f_2(X)].$$

A multiplicative separation is usually not possible. A discrete random variable X that is either 1 or 3 with a 50% chance has an expected value of $1\times 50\% + 3\times 50\% = 2$, while the expected value of X^2 is $1^2\times 50\% + 3^2\times 50\% = 5$. We see that, in general,

$$\mathbb{E}[X]^2 \neq \mathbb{E}[X^2].$$

A.3 Variance and Standard Deviation

The expected value expresses our average expectation of X. We'd also like to have a measure of a random's range of outcomes—its variability or volatility—around this expected value. For this, we examine $(X - \mathbb{E}[X])^2$—this expression becomes larger the more X tends to stray from its expected value. The average behavior of this expression is called *variance*, and it is defined as the following expected value:

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int (x - \mathbb{E}[X])^2\, p(x)\, dx.$$

We can use the properties of the expected value mentioned above to find an alternative expression for the variance as exercise (the expected value of an expected value $\mathbb{E}[\mathbb{E}[X]]$ is the constant $\mathbb{E}[X]$ inside):

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\
&= \mathbb{E}[X^2 - 2X\,\mathbb{E}[X] + \mathbb{E}[X]^2] \\
&= \mathbb{E}[X^2] - 2\,\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2.
\end{aligned}$$

We can use either expression to derive the following general properties:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[cX] &= c^2\,\mathbb{V}\mathrm{ar}[X], \\
\mathbb{V}\mathrm{ar}[c] &= 0.
\end{aligned}$$

The *standard deviation* is defined as the square root of the variance:

$$\mathrm{std}[X] = \sqrt{\mathbb{V}\mathrm{ar}[X]}.$$

It is often more useful than the variance because its scale or dimension is the same as that of X. If X values are in dollars, then the standard deviation lies on the same scale, while the variance has the unintuitive dimension of dollars-squared. (The variance's squaring approach merely helped make all the deviations from $\mathbb{E}[X]$ positively count toward our measure, and more gently so than the obnoxious absolute value.)

For our example of waiting times with its triangular probability density, we already know that

$$\mathbb{E}[X] = 2.$$

We compute

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2\, p(x)\, dx = \int_0^2 x^2\, (\frac{1}{4}x)\, dx + \int_2^4 x^2\, (1 - \frac{1}{4}x)\, dx = \cdots = \frac{14}{3}.$$

The variance becomes, via our shortcut,

$$\mathbb{V}\text{ar}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{14}{3} - 2^2 = \frac{2}{3}.$$

Its square root yields the standard deviation of 0.82 min or 49 s.

A.4 Sample Estimates

Funnily, not many of the concepts mentioned above are of immediate use—we usually do not know the probabilities p_i or the shape of the probability density function $p(x)$, and we therefore cannot compute the expected value or the variance. What we can do is make some observations and *estimate* them.

Recall our example of the waiting-time random variable, where we postulated a triangular probability density function that allowed us to compute the random's expected value. Instead of making such a sweeping assumption, we might also observe and record some actual waiting times, for example, {1:45, 0:23, 2:35, 3:17, 1:33, 2:10, 1:52}.

We call such sample observations x_i.[2] Given n such observations, we use the *sample mean*

$$\bar{x} = \frac{1}{n} \sum x_i$$

[2]This is fine for continuous distributions where these names are nowhere to be seen. Just make sure not to confound them with the outcomes of a discrete distribution.

and the *sample standard deviation*

$$s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$$

to estimate the real but unknown expected value and standard deviation.

For our example of waiting times, the sample above yields an estimate for the expected value of 1:56 (as opposed to the calculated 2 min) and one for the standard deviation of 54 s (close enough to the 49 s obtained theoretically)—all without those pesky integrals.

The estimator for the standard deviation warrants a few words. Our original definition of variance would, in the discrete, equiprobable case, translate into $\frac{1}{n}\sum(x_i - E[X])^2$—a prime suspect for an estimator. Why then use the unintuitive $n-1$ in our sample standard deviation?

There are several ways to frame an answer. Dividing by $n-1$ can be shown to yield an *unbiased* estimate for the variance, i.e., it doesn't err systematically, which sure is a welcome feature.[3] Most of the time we can get away with just using this unbiased variant. It's also the default way Excel's `STDEV` function operates.

Also, a market risk setup typically involves different statistical software packages, programming languages, and the odd Excel analysis; outside parties like regulators or consumers of risk reports might try to reenact the figures on their own systems. This—most commonly used—unbiased estimator ensures the desired exact comparability of results.

And finally, as the sample size n gets larger, the correction by -1 becomes ever less significant. A professor of mine once quipped that whenever you worry about this denominator, you really should be worried about your sample size.

But what is the mathematical rationale behind all this? Omitting theory, we can give an intuitive mnemonic aid. The variance is all about squared deviations from the expected value. Unfortunately, we don't know this expected value and have to estimate it via the sample mean. Assume, for example, two samples of $\{-2, +1, +2, -1\}$ and $\{+2, +1, +2, +1\}$ (of the same distribution), and notice how the entries in the latter, by chance, all point in the same direction. The first sample has a mean of $\bar{x} = 0$, which would have the variance estimator add up terms the like of 1^2 and 2^2. The second, somewhat more compact sample has, smack in the middle of its value range, a mean of $\bar{x} = 1.5$, which would have the variance estimator add up the smaller 0.5^2 terms. In fact, the sample mean always minimizes the sum of squared differences to itself, and because the unknown underlying expected value is usually different, this sum inevitably tends to undershoot the real variance. Luckily, it can be shown that the humble tweak of averaging the sum of squares over the smaller $n-1$ instead of n can swimmingly correct for this tendency.

[3] A minor detail for your next Jeopardy session: the standard deviation estimate is still not unbiased, due to the square root operation.

In some cases, we do not have to estimate the expected value via the sample mean because we know it—for example, if we impose our own when creating random values artificially, or if we have some knowledge or strong intuition about the underlying random's behavior. In such cases, we don't have to compute an $\bar{x}$ from the sample but can directly apply the knowledge of $E[X]$ in estimating the variance. It turns out that the *uncorrected* estimator (note the n instead of $n-1$ and the $\mathbb{E}[X]$ instead of $\bar{x}$) is then the way to go:

$$s = \sqrt{\frac{1}{n}\sum(x_i - \mathbb{E}[X])^2}.$$

We use this variant, for example, when illustrating the Monte Carlo modification in our VaR setup.

Some standard software packages support this directly, e.g., Python's `statistics.pvariance(data,mu)` function, which accounts for a known expected value. Often, however, implementations only use n in the denominator but still implicitly estimate the mean.[4]

A.5 Kurtosis

We have primarily looked at the average behavior of distributions and at their range or volatility of outcomes. This can often already give us a pretty good idea about a distribution. Observing, for example, the height of males, we might obtain a mean of roughly 172 cm and a standard deviation of about 7 cm. We can relate to these numbers: we know quite a few average-height people, some that are shorter or taller, and a select few that are extremely short or tall. We also certainly know very few people that are, for example, ten times the standard deviation of 7 cm (or 70 cm) *taller* than the average. All in all, we are confident of having a good grasp on the height range and might well be inclined to call its distribution "normal."

Now consider the number of Barbie dolls in households. This might often be 0, or 1, or 7, and maybe have a mean of 3 and a standard deviation of, say, 2 dolls. It is easy to imagine, however, that one avid collector in Wichita will own maybe 250 dolls (many, but too few to meaningfully impact the standard deviation itself). This is more than 100 times that standard deviation of 2, or a full 200 dolls, *above* the mean of 3. We didn't observe such strange behavior with heights—there, the same multiplier of the standard deviation would describe a giant, 7 m taller than the

[4]Many software packages default to the unbiased estimate (Matlab; Octave; S-plus; R; SAS; Mathematica; SPSS; Python's `np.cov` for calculating a covariance matrix).

Several implementations, by default, divide by n without accounting for the potentially known mean (Boost's `variance` function; Python's `np.var` and `np.std` functions.)

Often, alternative estimator functions are provided and can be used to coordinate disparate implementations. (Excel's `STDEV.P` divides by n; Python's `np.var` and `np.std` use $n-1$ when setting the optional argument `ddof=1`.)

average. The doll distribution, now, seems to exhibit such extreme outliers that are many standard deviations away from the average.

Distributions with such behavior are said to feature *heavy* or *fat tails.* To measure them, we need to smoke out such very large deviations. We achieve this by examining $(X-\mathbb{E}[X])^4$, whose hefty fourth power should bring them to our attention. Additionally, we'd like a measure of "tailedness" to also be independent of the scales or dimensions involved; the "number of limbs of Barbie dolls," about four times the original random number, should have the same heavy tail indicator as the doll distribution itself.

The following measure does this, and it is called *kurtosis*:

$$\mathbb{K}\text{urt}[X] = \frac{\mathbb{E}[(X - \mathbb{E}[X])^4]}{\mathbb{V}\text{ar}[X]^2}.$$

Normalizing by the variance in the denominator ensures our desired invariance under scaling:

$$\mathbb{K}\text{urt}[cX] = \mathbb{K}\text{urt}[X].$$

For reasons we will tackle soon, a good reference value for the kurtosis is 3. It indicates a benign tail behavior and no undue or extreme outliers. Larger kurtosis values indicate heavier tails and outliers more extreme than conventionally expected, and it is not uncommon to observe kurtosis values of 10 or even 50 in the wild. As for our examples:

- A million and one households, one with 250 Barbie dolls, 500,000 with 1 doll, and 500,000 with 5 dolls, have a combined and unsuspicious mean of 3.0002, a standard deviation estimate of 2.0152, but a whopping kurtosis of about 224.
- Our example of waiting times above has a kurtosis of 4.16, or nothing much to worry about.

As with the variance, we usually don't compute this kurtosis but instead estimate it from a sample. Many burdensome tweaks are required to obtain an unbiased estimate, as a quick Google search for "kurtosis estimator" will reveal. For our purposes, it suffices to rely on your preferred software package's implementation.

Be mindful of one thing, though: some kurtosis functions, like Excel's `KURT` one, report the so-called *excess* kurtosis, which is the kurtosis minus 3.

A.6 Multiple Random Variables and Covariance

So far we have examined an individual random variable X and its properties. We now take a look at how multiple randoms play together. Consider, first, two discrete random variables X and Y, where X can take on $x_1 = 0$ or $x_2 = 1$ and where Y can take on $y_1 = 0$, $y_2 = 3$, or $y_3 = 9$. Combined, we can obtain 6 different outcomes:

$(0,0)$, $(0,3)$, $(0,9)$, $(1,0)$, $(1,3)$, or $(1,9)$. Correspondingly, we need 6 probabilities (adding, again, up to 1) to describe the joint distribution, best expressed in a two-dimensional matrix:

$$\begin{pmatrix} p_{X=0,Y=0} & p_{X=0,Y=3} & p_{X=0,Y=9} \\ p_{X=1,Y=0} & p_{X=1,Y=3} & p_{X=1,Y=9} \end{pmatrix}.$$

We can proceed to naturally define the expected value of, for example, the sum of the two random variables over all outcomes:

$$\mathbb{E}[X+Y] = \sum_{i=1}^{2}\sum_{j=1}^{3}(x_i + y_i)\, p_{X=x_i,Y=y_j}.$$

(If we assume identical probabilities of $\frac{1}{6}$ for each outcome, we obtain a result of 4.5 for this expression.)

The expected value of X alone would be, again involving 6 terms,

$$\mathbb{E}[X] = \sum_{i=1}^{2}\sum_{j=1}^{3} x_i\, p_{X=x_i,Y=y_i}.$$

The one-dimensional approach we encountered at the beginning of this chapter can also be used, if we appropriately collect the involved probabilities:

$$\begin{aligned}\mathbb{E}[X] = 0 \times (&p_{X=0,Y=0} + p_{X=0,Y=3} + p_{X=0,Y=9}) \\ +1 \times (&p_{X=1,Y=0} + p_{X=1,Y=3} + p_{X=1,Y=9}).\end{aligned}$$

The two probabilities involved, each a sum of three of the original ones, express the events $X = 0$ and $X = 1$, irrespective of Y. Such "collapsed" probabilities are called *marginal* probabilities.

By analogy, we use two-dimensional density functions $p(x,y)$ and double integrals in case of continuous distributions:

$$\mathbb{E}[f(X,Y)] = \int\!\!\int f(x,y)\, p(x,y)\, dx\, dy.$$

Luckily, all this comes down to a simple conclusion—shuffle the terms around and convince yourself that we usefully always have

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[\,Y].$$

We can now begin to pose the first interesting question about multiple random variables: are they somehow related? For this, we examine whether they tend to move in the same direction. The expression $(X - \mathbb{E}[X])(Y - \mathbb{E}[\,Y])$ is positive if

both random variables are above their expected value, but it is also positive if both are below their expected value—this thus indicates co-movement in the same direction. The expression becomes negative, on the other hand, if they deviate in opposite directions from their expected values. Which of these types of co-movement dominates on average can be captured via the *covariance*:

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

If the covariance is positive, X and Y tend to move in the same direction on average; if it is negative, they tend to move in opposite directions. Either way, they two randoms are related.

We incidentally also note that

$$\mathbb{C}\text{ov}[X, X] = \mathbb{V}\text{ar}[X].$$

We shall mostly rely on estimators to guess the real but unknown covariance. In the special case where the individual expectations $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are known to be zero, the sample covariance estimate of n pairs of observations (x_1, y_1) to (x_n, y_n) simplifies to $\frac{1}{n}\sum x_i y_i$.

A convenient, normalized measure directly derived from this is the *correlation*, which yields values between $+1$ and -1, regardless of the volatilities underneath:

$$\text{corr}[X, Y] = \frac{\mathbb{C}\text{ov}[X, Y]}{\text{std}[X]\,\text{std}[Y]}.$$

(A quick intermediate sanity check: the variable X is surely strongly related to itself, as X always moves in the same direction as X (d'oh). If we actually evaluate the correlation of X to itself, we get $\text{corr}[X, X] = 1$. Likewise, the correlation of X to its opposite $-X$ is $\text{corr}[X, -X] = -1$.)

The covariance only approximates how random variables interact. For a deeper apprehension, we need the concept of *dependence*. It can be approached as follows: if knowing the outcome of one random variable does not give you any hint or additional information on how the other random variable will behave, then the two randoms are called *independent*.

It turns out that independent randoms always have a covariance or correlation of zero, and non-zero covariance or correlation thus signals dependence. Let's wrap our heads around this. If, for example, two randoms are positively correlated and we know that the first one went up, we'd expect the second random to tend to do the same. This is definitely tangible information, and it follows that the two randoms can't be independent.

When meeting a statistician at a bar, it is useful to keep in mind that zero correlation does not, in turn, guarantee independence. Witness the two randoms X (uniform in $[-1, 1]$) and $Y = X^2$. They are clearly dependent, for knowing the

outcome of X will already foretell us the exact outcome of Y, but they have zero correlation. Yet such instances are rare in our context, and you'll find that zero correlation will many times correctly hint at independence.

We are now ready to tackle the issue of the variance of random sums. We already know that $\mathbb{V}\text{ar}[X - X] = 0$ and that $\mathbb{V}\text{ar}[X + X] = 4\,\mathbb{V}\text{ar}[X]$—a hint, maybe, that the relation between the random numbers might affect the variance of their sum. But let's plow through:

$$\begin{aligned}
\mathbb{V}\text{ar}[X + Y] &= \mathbb{E}[((X + Y) - \mathbb{E}[X + Y])^2] \\
&= \mathbb{E}[(X + Y)^2 - 2(X + Y)\,\mathbb{E}[X + Y] + \mathbb{E}[X + Y]^2] \\
&= \mathbb{E}[(X + Y)^2] - \mathbb{E}[X + Y]^2 \\
&= \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X] + \mathbb{E}[\,Y])^2 \\
&= \mathbb{E}[X^2] + 2\,\mathbb{E}[XY] + \mathbb{E}[\,Y^2] - \mathbb{E}[X]^2 - 2\,\mathbb{E}[X]\,\mathbb{E}[\,Y] - \mathbb{E}[\,Y]^2.
\end{aligned}$$

If we look carefully at the last line's terms, we find that

$$\mathbb{V}\text{ar}[X + Y] = \mathbb{V}\text{ar}[X] + \mathbb{V}\text{ar}[\,Y] + 2\,\mathbb{C}\text{ov}[X, Y].$$

We get an even nicer expression for *independent* randoms, whose covariance, as mentioned above, is zero:

$$\mathbb{V}\text{ar}[X + Y] = \mathbb{V}\text{ar}[X] + \mathbb{V}\text{ar}[\,Y].$$

Going one step further, we can determine the variance of sums of more than two randoms. The number of terms becomes a bit unwieldy, but, fortunately, we can express the final result in a conveniently brief matrix notation. With a row vector of constants $\mathbf{c}$, recalling that $\mathbb{V}\text{ar}[X] = \mathbb{C}\text{ov}[X, X]$, and defining a covariance matrix $[\mathbf{C}]$ with entries $\mathbb{C}\text{ov}[X_i, X_j]$, we conclude, with some patience, that

$$\mathbb{V}\text{ar}[c_1X_1 + c_2X_2 + c_3X_3] = \mathbf{c}[\mathbf{C}]\mathbf{c}^\top.$$

A.7 Distribution, Inverse, and Quantiles

The probability that $X < q$ is given by

$$\int_{-\infty}^{q} p(x)\,dx.$$

This recurring concept is best abbreviated via the *cumulative distribution function* $P(x)$:

$$P(x) = \int_{-\infty}^{x} p(t)\,dt.$$

The probability that $X < q$ is then $P(q)$.

The probability that X lies in a certain range $[q_d, q_u]$ is

$$\int_{-\infty}^{q_u} p(x)\,dx - \int_{-\infty}^{q_d} p(x)\,dx = P(q_u) - P(q_d).$$

The probability that $X > q$ is $1 - P(q)$.

We already know that the probability of $X = q$, i.e., of X ending up in the zero-length interval $[q, q]$, is $P(q) - P(q) = 0$. This odd property, once it has been shruggingly accepted, has the nice consequence that we need to worry less about open/closed intervals or the difference between" $<$" and "$\leqslant$"—the infinitesimally small "border" outcomes make (for practical intents and purposes) no difference. This also makes many expressions for continuous distributions simpler, whereas, for discrete ones, we have to be much more careful about indices at boundaries.

Now let's do the reverse: given a probability p, we can use the inverse $P^{-1}(\cdot)$ to find the corresponding value q such that $P(q) = p$:

$$q = P^{-1}(p).$$

Such q-values are called *quantiles*. The 1%-quantile $q_{1\%} = P^{-1}(1\%)$, for example, is our 1%-value-at-risk.

It makes sense to name or index the quantiles with their corresponding probabilities. The probability of a random number falling between $q_{3\%}$ and $q_{7\%}$, for instance, then becomes

$$P(q_{7\%}) - P(q_{3\%}) = 7\% - 3\% = 4\%.$$

What are the 5% shortest waiting times in our triangular waiting time distribution, i.e., what is its 5%-quantile? For quantiles q on the left side of the triangle, the integral in the cumulative is simply the triangular area $P(q) = \frac{1}{2} q \frac{q}{4} = \frac{1}{8} q^2$. For a probability $p = \frac{1}{8} q^2$, the inverse becomes $q = P^{-1}(p) = \sqrt{8p}$. Thus, for $p = 5\%$, the quantile $q_{5\%} = \sqrt{8 \times 5\%} = \sqrt{4/10}$. Verifying this, we see that indeed $P(q_{5\%}) = \frac{1}{8} q_{5\%}^2 = \frac{4}{80} = 5\%$. The 5% shortest waiting times lie between 0 and 38 s.

Quantiles, just like the expected value, scale and translate under linear transformations of the type $Y = aX + b$, with $a > 0$. You can formally prove this, or you can consider this to be simply a change of measurement units, like transforming Celsius to Fahrenheit, and sign off on it. The distribution's core characteristics remain, and only the involved dimensions change. We have, e.g., for the 1%-quantile,

$$q_{1\%}^{Y} = a\, q_{1\%}^{X} + b.$$

A.8 Conditional Expectation

Sometimes only a subset of outcomes is of interest to us. The expected value, for instance, of a random variable under such a restricting condition is called *conditional expected value*. A typical example is the expected value of *X if X is smaller than a certain number c*. To obtain it, we simply sum/integrate up only to that number and normalize the result with the probability of our condition:

$$\mathbb{E}[X|X < c] = \frac{1}{P(c)} \int_{-\infty}^{c} x\, p(x)\, dx.$$

Our risk measure of the expected shortfall is such a conditional expectation. It deals with the 2.5% largest losses, so we have $c = q_{2.5\%}$ and $P(c) = P(q_{2.5\%}) = 2.5\%$. In our discrete case, we sum up 25 values of interest (the largest losses), each weighted with a probability $p_i = \frac{1}{1000}$. Dividing by the overall probability of 2.5% leaves us with the denominator 25 in Eq. (8.1).

A.9 The Normal Distribution

We finally get to meet an important and ubiquitous kind of distribution, one so common as to be called *normal* distribution. It arises in the context of sums of random variables, by which many phenomena can be characterized. A leaf falling through the air, for example, will undergo a series of tiny random nudges hither and thither before hitting the ground. A heap of leaves below a tree is then normally distributed. But onwards, from the bucolic to the more prosaic.

The normal distribution's probability density function $p(x)$ is driven by two parameters, μ and σ:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

This is denoted as an $N(\mu, \sigma^2)$-distribution. It is shaped like a bell, as can be seen in Fig. A.2, and also called a bell curve or a Gaussian.

This particular parameter setup is chosen to make the integral expressions for the base measures conveniently evaluate to

$$\begin{aligned}
\mathbb{E}[X] &= \int x\, p(x)\, dx = \cdots = \mu, \\
\mathbb{V}\mathrm{ar}[X] &= \int (x - \mathbb{E}[X])^2\, p(x)\, dx = \cdots = \sigma^2, \\
\mathrm{std}[X] &= \sigma.
\end{aligned}$$

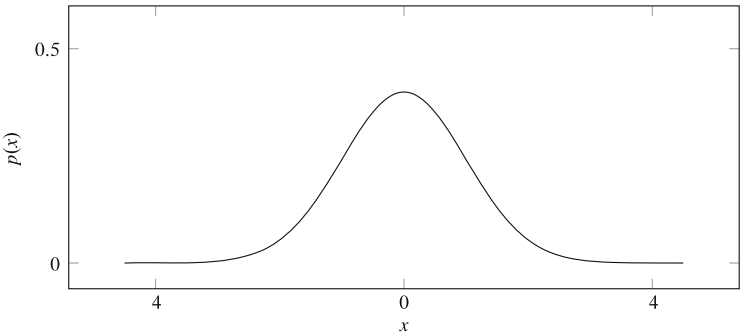


Fig. A.2 Normal probability density, for $\mu = 0$ and $\sigma = 1$

In theory, a normal random can take on any value—notice how the density is positive for all real numbers x, which allows for arbitrarily large positive or negative outcomes to occur. In practice, the probabilities become small so fast (due to the mighty e^{-x^2} term) that extreme events far away from μ are highly unlikely. The tails of a normal are thus not heavy but rather ordinary. The normal's kurtosis in fact came to signify unexciting and boring tail behavior. For any normal, regardless of its standard deviation, we have

$$\mathbb{K}\text{urt}[X] = 3.$$

That's where our ominous kurtosis value of 3 in Sect. A.5 originates.

A normal's cumulative and its inverse have no closed-form solution; we evaluate them by numerical approximation or by referring to tables with pre-computed values. For a normal distribution with $\mu = 0$ and $\sigma = 1$ (i.e., an $N(0, 1)$-distribution, also called *standard normal* distribution), the cumulative distribution is usually called Φ. Its 1%-quantile is approximately

$$q_{1\%} = \Phi^{-1}(1\%) = -2.32635.. \approx -2.33.$$

Quantiles scale with σ.[5] For an $N(0, \sigma^2)$-normal, we have

$$q_{1\%} = P^{-1}(1\%) = \sigma \times \Phi^{-1}(1\%) \approx -2.33\,\sigma.$$

To calculate the constant, use Excel's `NORM.INV(0.01;0;1)` or Wolfram Alpha's `InverseCDF[NormalDistribution[0,1], 0.01]`.

[5]That's because it can be proven that each $N(\mu, \sigma)$-normal can be expressed as a scaled $\sigma X + \mu$, with X standard normal. Since quantiles scale in general under such linear transformations, all normal quantiles can be retraced back to the standard ones, and we don't have to explicitly recalculate them for each different parameter combination.

The expected shortfall of a normal is the following conditional expectation:

$$\begin{aligned}\mathrm{ES}[X] &= \mathbb{E}[X|X < q_{2.5\%}] \\ &= \frac{1}{P(q_{2.5\%})} \int_{-\infty}^{q_{2.5\%}} x\, p(x)\, dx \\ &= \frac{1}{2.5\%} \int_{-\infty}^{q_{2.5\%}} x\, p(x)\, dx.\end{aligned}$$

For a standard normal with a density function φ, this evaluates to

$$\mathrm{ES}[X] = -\frac{1}{2.5\%}\varphi(\Phi^{-1}(2.5\%)),$$

and it is approximated numerically as 2.33780, e.g., using Excel's

```
-NORM.S.DIST(NORM.S.INV(0.025);FALSE)/0.025
```

It also scales with σ.

A.10 Sums of Randoms

We have already encountered random sums like $X + Y$. Still, we now get our hands dirty and try to become a bit more acquainted with them. Warning: your best friends in this section will be a pen and some sheets of paper.

Consider two independent random numbers with a uniform distribution, say, X over the interval $[0, 2]$ and Y over the interval $[0, 3]$. Their two-dimensional probability density function is

$$p(x, y) = \begin{cases} \frac{1}{6} & \text{for } x \text{ in } [0, 2] \text{ and for } y \text{ in } [0, 3], \\ 0 & \text{otherwise.} \end{cases}$$

Each random has, individually, its own (one-dimensional) probability density function. As its area must sum up to 1, we must have

$$p^X(x) = \begin{cases} \frac{1}{2} & \text{for } x \text{ in } [0, 2], \\ 0 & \text{otherwise;} \end{cases}$$

$$p^Y(y) = \begin{cases} \frac{1}{3} & \text{for } y \text{ in } [0, 3], \\ 0 & \text{otherwise.} \end{cases}$$

The relation between $p(x, y)$ and $p^X(x)$ or $p^Y(y)$ goes deeper. Just like we "collapsed" 6 discrete probabilities into 2 marginal ones in Sect. A.6, we may also determine $p^X(x)$ (for x in $[0, 2]$; it is zero otherwise) by integrating over y:

$$p^X(x) = \int_{-\infty}^{\infty} p(x, y)\, dy = \int_0^3 p(x, y)\, dy = \int_0^3 \frac{1}{6}\, dy = \frac{1}{2}.$$

Equivalently, integrating over the x-dimension yields the Y-marginal p^Y.

A very similar collapse from two dimensions to one will help us tackle the probabilities of random sums. If we define a new random variable $Z = X + Y$, with X and Y uniform as above, we might ask: What is Z's one-dimensional probability density? Is Z maybe also uniformly distributed?

To tackle this, we briefly digress to the discrete, two-dimensional setup of a uniformly random chess board with discrete axes X and Y between 1 and 8. The probabilities of the 64 outcomes $X = i$ and $Y = j$ are $p_{X=i,Y=j} = \frac{1}{64}$. What about a random $Z = X + Y$ and its (one-dimensional) probabilities $p_{Z=k}$?

First, $Z = 1$ can never happen, as it will always at least be 2.

There is only one way our Z can become 2: if both X and Y are 1 (all other setups create a larger Z). The corresponding probability is thus

$$p_{Z=2} = p_{X=1,Y=1}.$$

There are two ways to obtain $Z = 3$: via $X = 1$ and $Y = 2$, or via $X = 2$ and $Y = 1$. The corresponding probability is

$$p_{Z=3} = p_{X=1,Y=2} + p_{X=2,Y=1}.$$

There are three ways to obtain $Z = 4$: via $X = 1$ and $Y = 3$, via $X = 2$ and $Y = 2$, or via $X = 3$ and $Y = 1$:

$$p_{Z=4} = p_{X=1,Y=3} + p_{X=2,Y=2} + p_{X=3,Y=1}.$$

You get the idea—we basically sum up probabilities over diagonal segments of our board to obtain the $p_{Z=k}$. And it is also clear that these probabilities differ—that Z is not uniform. Its probabilities are as follows:

- The white main diagonal corresponds to $p_{Z=9}$:

$$p_{Z=9} = \sum_{t=1}^{8} p_{X=t,Y=9-t} = \frac{8}{64}.$$

- The lower left diagonals yield 7 probabilities for $Z = k$, with k between 2 and 8:

$$p_{Z=k} = \sum_{t=1}^{k-1} p_{X=t,Y=k-t} = \frac{k-1}{64}.$$

- The upper right diagonals also yield 7 probabilities for $Z = k$, with k between 10 and 16:

$$p_{Z=k} = \sum_{t=k-8}^{8} p_{X=t,Y=k-t} = \frac{17-k}{64}.$$

With a good hunch, we return to our continuous two-dimensional distribution. Because the original $[0, 2] \times [0, 3]$-uniform is a bit tedious with regard to integration bounds, we consider the simpler uniform on $[0, 1] \times [0, 1]$ with $p(x, y) = 1$ over that area. We confidently declare that the following density describes $Z = X + Y$:

$$\begin{aligned} p^Z(z) &= \int_{-\infty}^{\infty} p(t, z-t)\, dt \\ &= \begin{cases} \int_0^z p(t, z-t)\, dt = t|_0^z = z & \text{for } z \text{ in } [0, 1], \\ \int_{z-1}^1 p(t, z-t)\, dt = t|_{z-1}^1 = 2 - z & \text{for } z \text{ in } [1, 2], \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The probability density function of Z is thus a triangle. (As an exercise, you might want to try this for uniforms of unequal ranges.)

There are several reasons we went through this exercise. It should, first, underline the close correspondence of discrete and continuous setups. It also hopefully illustrates that boundary cases can often be managed more elegantly, and with less of an index mess, in a continuous setup. Mainly, however, is should stress that probability densities of randoms do not always translate trivially into the density of their sum. This should prepare the stage for what hopefully provides some relief now.

For it turns out that multiple *normal* random variables following the so-called *multi-variate* normal distribution behave much more benignly under summation. Each random variable is normally distributed, and, crucially, it can be shown that their sum is also normally distributed, which spares us laborious integrals. We can derive the characteristics (i.e., μ and σ) of the sum of normals directly from the individual distributions' μ, σ, and their correlation ρ:

$$\begin{aligned} \mu_{X+Y} &= \mu_X + \mu_Y, \\ \sigma_{X+Y} &= \sqrt{\sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y}. \end{aligned}$$

The main takeaway, neglecting normals that follow degenerate "non-multi-variate" distributions, is that "the sum of normals is normal"—if for no other reason than because we often simply assume such a distribution off the bat.[6]

[6]Two more things to keep in mind: two independent normals also share a joint density, and the expressions above simplify further because their $\rho = 0$; and for multi-variate normals, independence and zero correlation are fully equivalent concepts.

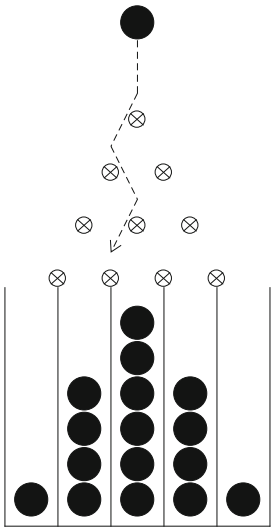


Fig. A.3 A Galton board

There are several ways to prove this, but they often only provide verification in a formal, technical sense. Surely such a fundamental property must be rooted understandably in the very setup of the normal distribution itself. So to instill some confidence, we look at the case of independent normals and their sum via the so-called Galton board depicted in Fig. A.3.

In this game, a ball is dropped over layers of offset nails. As it traverses downwards, it randomly goes left or right at each nail, before finally ending up in a bin below. The height of the ball stacks in the bins—the outcome of random sequences or sums of left/right movements—can be shown to resemble a normal density as we use more and more layers of nails. Assume that one such board corresponds to a normal X. Now, let's drill a hole in a bin below, attach a second board right below that hole, let the balls fall on, and collect them again further down; we repeat this for each bin. This is akin to adding a second (also normal) board Y to the first one. The whole procedure should result in the same final bin tally as when using one larger board with as many layers as X and Y combined. Because such a larger board is, like any board, also akin to a normal, the sum of the original boards better be as well.

We can use a similar trick with our sum of uniforms, whose probability density we already found to be triangular. This time the game is Tetris. We let a first random X in $[0, 4]$ determine the starting position of the coveted 1×4 brick. Each such brick we then interpret as the probability density of width 4 of a second uniform Y, right before we let it drop down. Once such a brick, starting off at position X, comes to rest, we consider it to stand for part of the density of $X + Y$. As each starting position X is equally likely, we might as well loop through them, obtaining the left-hand side in Fig. A.4.

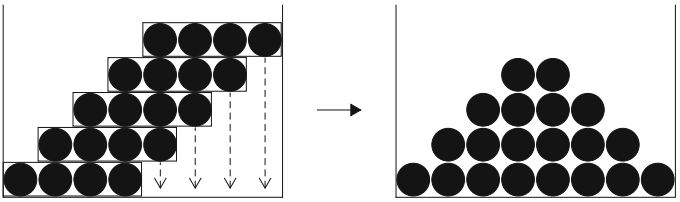


Fig. A.4 Tetris

If we then let the bricks unglue and the resulting 1×1 pieces drop to their final resting place, we obtain the $X + Y$ density—our familiar triangle.

A.11 Some Densities for the Road (to Independence)

We have seen that we were able to derive quite a few properties of randoms without much actual computation. Some distributions, most notably the normal one, provide additional shortcuts because of their very specific structure. Still, it is often instructive to perform a handful of raw calculations explicitly to whet our intuition, especially in the two-dimensional case. Here are some starting points.

We have already encountered the triangular distribution in our waiting time example. What would a two-dimensional probability density of two independent waiting times X and Y over $[0, 4] \times [0, 4]$ look like? To get an idea, go to your bedroom, grab your bed sheet right at the center of the mattress, and pull it up. The resulting structure resembles a wigwam. Let's try to construct a corresponding probability density $p(x, y)$.

The wigwam has its peak at $(x, y) = (2, 2)$. Let's look at the lower left part or quarter of this volume first (i.e., x in $[0, 2]$ and y in $[0, 2]$). If we define $p(x, y) = cxy$, we see that its height is 0 for $x = 0$ or $y = 0$, and it is $4c$ for $(x, y) = (2, 2)$. The volume of the lower left part is[7]

$$\int_0^2 \int_0^2 p(x, y)\, dx\, dy = \int_0^2 \int_0^2 cxy\, dx\, dy = 4c.$$

To also describe the remaining 3 quarters of the wigwam, we may use, for x and y both in $[0, 4]$,[8]

$$p(x, y) = c(2 - |2 - x|)(2 - |2 - y|).$$

[7] You can use Wolfram Alpha's website and type in
`integrate c x y dx dy, x = 0 to 2, y = 0 to 2`
to make sure.

[8] You can directly google `(2-abs(2-x))(2-abs(2-y))`, which should give you a nice 3D-plot where you just need to adjust the graph's display ranges.

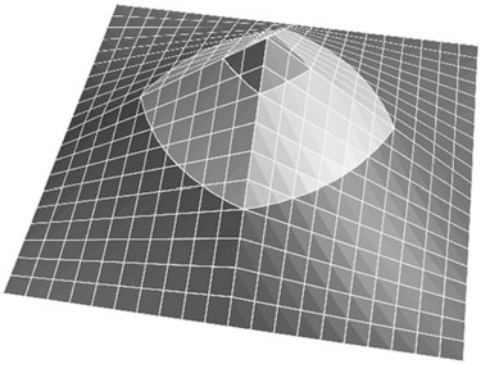


Fig. A.5 Wigwam (z-axis range tuned for clarity)

The total volume of the wigwam must be, because of its symmetry, $4 \times 4c = 16c$. For $c = 1/16$, we obtain a valid probability density (see Fig. A.5).

Let's check the marginal distribution of X, for x in $[0, 2]$ (larger x work similarly):

$$p^X(x) = \int_0^2 \frac{1}{16} xy \, dy + \int_2^4 \frac{1}{16} x(2 - (y - 2)) \, dy = \cdots = \frac{1}{4} x.$$

This is the left side of our trustworthy triangle distribution. For the full range of x in $[0, 4]$, we get

$$p^X(x) = \frac{1}{4}(2 - |2 - x|).$$

We notice that in our wigwam case we have $p(x, y) = p^X(x)p^Y(y)$—so we could have avoided all the construction work and simply have multiplied the individual densities in the first place. Such a neat multiplicative separation of a two-dimensional probability density is not always possible. If it is, though, then this is very telling, as we will soon discover.

We are almost done with our wigwam but still want to check whether the X and Y described by it are independent. Their covariance is zero (as you can verify by doing the appropriate integration), but that's only a hint. Recall that independence essentially means that knowing X does not tell us anything about Y.

How can Y behave if we know that X equals some specific, say, x'? We intuit that it should loosely behave according to the one-dimensional function of y given by $p(x', y)$, i.e., a vertical slice through our wigwam. This slice is always a triangle here, as depicted in Fig. A.6.

This almost looks like a density already, except that its area does not have to be 1. We can easily remedy that by scaling the function and dividing it by its own area $\int p(x', y) \, dy$, which of course is simply the value given by $p^X(x')$. Doing this yields a valid *conditional* probability density with area 1:

$$p(y|X = x') = \frac{p(x', y)}{p^X(x')}.$$

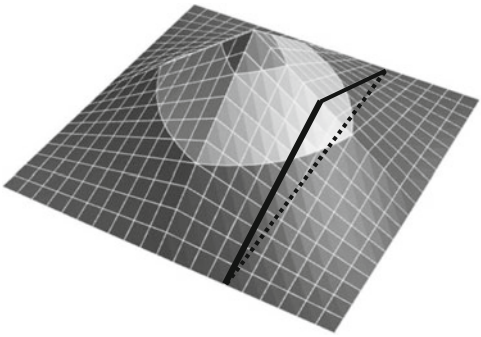


Fig. A.6 Wigwam with (yet unscaled) conditional density

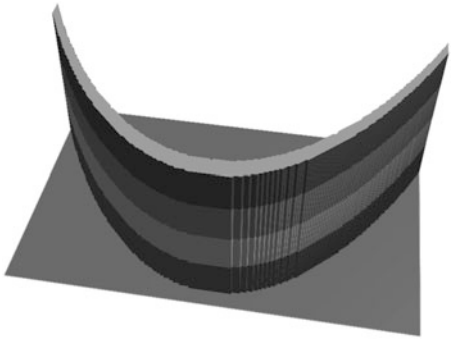


Fig. A.7 Plank

This density describes how Y behaves once X has settled on x'. The question of independence becomes: does this x' even influence Y? At first sight, yes (there are, after all, plenty of x' on the right-hand side of the equation). At second sight, we recall that for the wigwam it holds that $p(x, y) = p^X(x)p^Y(y)$, and therefore

$$p(y|X = x') = \frac{p(x', y)}{p^X(x')} = \frac{p^X(x')p^Y(y)}{p^X(x')} = p^Y(y).$$

So Y's conditional density is not affected by X at all—a realization of X does not tell us anything about how Y might behave. The wigwam must be independent. (It is also but a little mental stretch that makes us realize: independence and the multiplicative separation of a two-dimensional density mean one and the same.)

Now take two randoms, with X uniform in $[-1, 1]$ and Y uniform in $[x^2, x^2+0.01]$. The graph looks like a bended plank standing upright on its narrow side, 2 wide and 0.01 thick (see Fig. A.7). How long is it, i.e., how high is the graph? Well, the volume must be $1 = 0.01 \times 2 \times h$, so we have a height of 50.

The conditional density of Y given $X = x'$ is

$$p(y|X = x') = \begin{cases} \frac{50}{0.01 \times 50} = 100 & \text{for } y \text{ in } [x'^2, x'^2 + 0.01], \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, we can't get rid of the x' here. Knowing x' is in fact vital and, in turn, also inevitably tells us a lot about how Y will behave. If X is close to 1 or -1, then

Fig. A.8 Pyramid

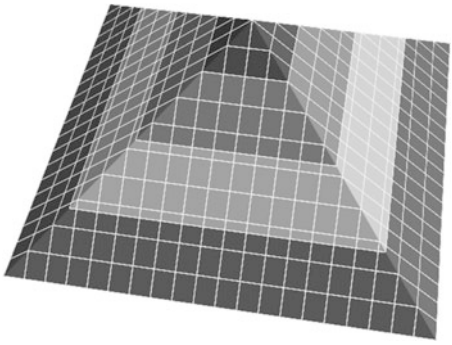
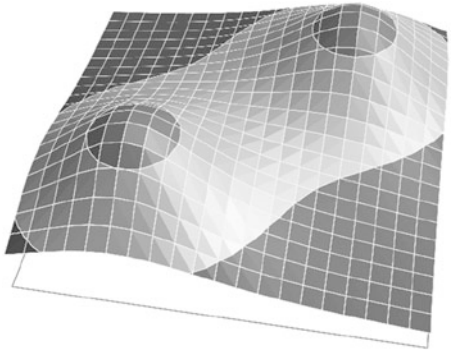


Fig. A.9 Camel



Y will be close to 1; if X is close to zero, so will be Y. These X and Y are therefore not independent.[9]

With this intuition, we can make short shrift of another density, the pyramid density of Fig. A.8 over the base $[0, 4] \times [0, 4]$, peaking at $(2, 2)$. Are the thusly described X and Y independent? Well, the conditional densities around values of x' close to 2 look like triangle distributions, whereas the ones around $x' = 1$ or $x' = 3$ look trapezoid. No amount of mere scaling can ever bring them in line—Y behaves differently for different x'. The pyramid is not independent.

To wrap up, lest we get the impression that there are only freak distributions out there, it helps to construct a plain one with a proper covariance. We might call this one a camel hump distribution. To obtain peaks at, say, $(1, 1)$ and $(3, 3)$, we start with the following guess[10]:

$$p^?(x, y) = \frac{1}{1 + (1 - x)^2 + (1 - y)^2} + \frac{1}{1 + (3 - x)^2 + (3 - y)^2}.$$

[9]Their covariance, though, is zero, as a calculation exercise reveals. Also note that the already encountered randoms X and $Y = X^2$ are the limiting case of ever-thinner planks.

[10]Google `1/(1+(1-x)^2+(1-y)^2) + 1/(1+(3-x)^2+(3-y)^2)` to confirm this function's "camelity," or refer to Fig. A.9.

Its volume on $[0,4]\times[0,4]$ is 9.4774,[11] which normalizes our guess into a proper density $p(x,y) = \frac{1}{9.4774}p^?(x,y)$. The resulting covariance is, finally, a full-fledged number:

$$\mathbb{C}\mathrm{ov}[X,Y] = \int\int (x-2)(y-2)p(x,y)\,dx\,dy = \cdots = 0.342146.$$

[11] Via Wolfram Alpha's
`integrate (1/(1+((1-x)^2+(1-y)^2))+1/(1+((3-x)^2+(3-y)^2))) dx dy, x=0 to 4, y=0 to 4.`

Pricing

B

Determining the value of your positions is easy for frequently traded, *liquid* assets because market quotes of prices are readily available. But what about completely new, never-before traded assets without market quotes, or assets that are very rarely traded, so-called *illiquid* assets? Determining their value cannot purely rely on direct lookups into a current market snapshot, as there are no or only sporadic records of actual trades.

The idea in pricing such positions is not spectacular and relies on one fundamental assumption: *asset prices should be consistent*. For example, bonds with similar maturities and from issuers of similar credit-worthiness should probably cost about the same. Likewise, currency exchange rates should be attuned to each other. If given two exchange rates $\$_{€}$ and $€_{£}$, then the direct exchange rate $\$_{£}$ should be in line with them. Otherwise there would be a cheap way to convert **$** into **£** (for example, by converting **$** into **€** and then **€** into **£**) and a more expensive one (the other route of directly converting **$** into **£**), and traders taking the cheap route would bid up the prices involved until both ways of conversion aligned again.

Following this train of thought, another way of expressing this idea of price consistency comes to mind: there should be no sure, risk-free profits. If you knew that a stock is currently worth 12**$** and someone offered it to you for 10**$**, you could buy it and immediately sell it on the market—pocketing a sweet, risk-free 2**$** in the process. Price consistency means that such discrepancies are assumed not to exist. Gains like these would be the financial world's equivalent of the physical impossibility of doing work without spending energy.

This reasoning can even be applied if it involves the fog of the future: you should not be able to schedule prospective asset exchanges in a way that guarantees you a risk-free profit. Assume, for example, that a stock **S** is currently worth 12**$**. Someone offers you a deal: if you pay him 10**$** now, he will deliver this stock in one month's time. It seems an OK deal, and you could accept it, pay, wait a month, and hope the stock will later be worth more than what you paid now—yet you still can both gain or lose, i.e., you take on some risk. However, you could transform this deal into

M. Auer, *Hands-On Value-at-Risk and Expected Shortfall*, Management for Professionals, https://doi.org/10.1007/978-3-319-72320-4

one with guaranteed risk-free profit the following way: borrow the stock, sell it for its current price of 12$, pay 10$ for the original deal, wait a month, get the stock as promised, and return it to the stock lender. You are left with a sure profit of 2$, regardless of how the stock develops. Assuming no risk-free profits exist thus rules out such deal opportunities as well.

Extending this one step further, we even want to ban risk-free profits *on average*, for example, for repeated, risky deals. Say, someone offered to (repeatedly) cast a die and to give you the resulting number of stock units—how much would you be willing to pay to take part in this game? If he offered you this deal in exchange for three units of stock, you would certainly take it and just keep playing, for on average you'd get 3.5 units of stock by investing only 3. As above, it seems reasonable to rule out such deals as well.

In short, exchanging assets back and forth should not let you end up with more asset units than you started out with. Such a profit bonanza is called *arbitrage* opportunity (there is a precise mathematical definition of arbitrage, but we'll keep it shamelessly visceral). Finding consistent prices relies on the assumption that no such arbitrage exists, and it is called *no-arbitrage pricing*. Another angle on this is that prices of illiquid assets are *interpolated* from known prices of liquid ones, and that the discipline of pricing, known as *quantitative finance*, is actually a big interpolation framework.

But wait—didn't we already encounter an arbitrage opportunity? We could buy a zero bond for 0.9$ and thereby make sure to earn 0.1$ once the bond pays out its promised 1$. This seems to be a sure profit—but only on the face of it. Money, unfortunately, usually loses value over time, and comparing nominal units of money at different points in time is therefore misleading. A can of Coke, for example, cost 5 cents in the 1950s; it costs more now but is probably "worth" the same as back in the day.

So money is a special kind of asset, also called (negative) arbitrage asset for its holder. Like cars, money typically loses value over time with respect to other assets. It is a (positive) arbitrage asset for the issuer of the money, the government. This makes—ironically—monetary prices, i.e., prices with respect to money, somewhat ill-suited for consistent pricing. We shall now get to know a more elegant, money-eschewing approach to arbitrage-free pricing. We can only hint at its most basic ideas here. But after some simple examples, we should at least be able to price call options, which was worthy of a cool Nobel prize not too long ago.

Note: this brief motivational chapter heavily borrows both ideas and notations from Jan Vecer's "Stochastic Finance: A Numeraire Approach" (Vecer 2011).

B.1 Trades as Asset Exchanges

Financial assets are often described as being intangible, but I find that it sometimes helps to view them as existing, palpable, and immutable things. Like a dollar bill, we can consider a bond or a stock to be a piece of paper or a contract (and not so long ago, before online banking, those papers actually existed). In this chapter, we will

denote assets in bold face—a dollar **$**, a stock **S**, a zero bond **B**, etc., to demarcate assets from their prices. Assets do not change over time—a piece of paper remains a piece of paper. Their prices, however, do change.

Positions and portfolios are then quite naturally mere multiples and sums of such assets, e.g., a portfolio might consist of some stock, bonds, and debt:

$$80\mathbf{S} + 400\mathbf{B} - 1200\$.$$

If we borrow one unit of a stock and sell it to bet on falling prices, i.e., if we short it, the resulting portfolio is

$$-\mathbf{S} + 12\$.$$

This way of describing positions and portfolios is handy for keeping track of rights and obligations. It can also express asset trades and deal with time. We can, for example, describe buying a stock now (at time $t = 0$) with the following asset exchange relation:

$$\mathbf{S} \sim^0 12\$.$$

This asset relation denotes that 12**$** can (now) be traded for one unit of **S**, or vice-versa. The stock's price in **$** is 12. We can also express promises this way, for example, that a zero bond **B** will pay 1**$** at time T corresponds to the following future exchange:

$$\mathbf{B} \sim^T \$.$$

A contract **F** that obliges you to buy a stock at a set dollar price k in the future is given by

$$\mathbf{F} \sim^T \mathbf{S} - k\$.$$

Think of it as exchanging, at time T, a piece of paper called **F** for a piece of paper called **S** while parting with k precious pieces of paper called dollars.

The math of asset relations behaves intuitively. Adding or subtracting assets and grouping together assets of the same kind make sense; operations like multiplying an asset with another do not. Note that asset relations are valid only at a specific point in time, so even if it holds that $\mathbf{B} \sim^T \$$, it does not follow that a bond can be exchanged for a dollar right now ($\mathbf{B} \not\sim^0 \$$). A zero bond is typically cheaper than the future dollar it promises; we might, e.g., experience current exchange levels of $\mathbf{B} \sim^0 0.9\$$. Expressed in terms of some continuously compounded interest rate r, we often equivalently express this as $\mathbf{B} \sim^0 e^{-rT}\$$.

B.2 Prices as Ratios

Asset prices are exchange ratios that describe how many units of an asset can be exchanged for a unit of a different asset. Buying a stock **S** worth 12**$** can be expressed, as we have seen, in the following asset exchange relation:

$$\mathbf{S} \sim^0 12\$.$$

Another way to describe this is to say "the price of **S** now, at time 0, in terms of **$** is 12" or, less chatty,

$$S_\$(0) = 12.$$

This is no longer an asset relation but a conventional mathematical equation of prices or ratios. The font face alerts us: $S_\$(0)$ is a number, while **S** is a thing.

The so-called reference asset used for pricing in the example above is the dollar **$**. But we can also express price ratios with respect to another asset, maybe a zero bond **B**. The price of **S** with respect to **B** is the number of units of **B** needed to buy one unit of **S**. How to get this new price? We know that we can exchange $\mathbf{S} \sim^0 12\$$. If the current bond price is $B_\$(0) = 0.9$, we can exchange $\mathbf{B} \sim^0 0.9\$$ or $\$ \sim^0 \frac{1}{0.9}\mathbf{B}$. So we can exchange

$$\mathbf{S} \sim^0 12\$ \sim^0 12 \times \frac{1}{0.9}\mathbf{B} \sim^0 13.33\mathbf{B}.$$

We thus obtain the current price of **S** with respect to **B**:

$$S_B(0) = 13.33.$$

Why would we ever want to use reference assets other than the **$**? The answer is that some pricing exercises become simpler. By sidestepping money as reference assets, we can often avoid having to compensate for its depreciation via discounting or its opposite, compounding. In more complex setups, we might be able to reduce the dimensionality of integrals. In short, it is simply more elegant.

So we mainly operate on prices with respect to no-arbitrage assets and in the end convert those prices to dollar ones via chained relations like

$$S_\$(0) = S_B(0)B_\$(0).$$

B.3 Prices of Future Delivery

The dollar prices of liquid assets are given by the current exchange ratios readily visible in the markets. But what happens if we want to exchange assets in the future? After all, asset prices fluctuate and the future is unknown.

This section deals with two such examples and involves, even though future prices are random, no probabilities. Both rely on the idea of no-arbitrage. The first answers how much you should be willing to pay *now* in order to get one unit of **S** in the future. The second is about how much you should agree to pay *in the future* for that same **S**.

Consider a contract **K** that promises to deliver, at time T, one unit of **S**:

$$\mathbf{K} \sim^T \mathbf{S}.$$

What is this contract worth now, i.e., what is its price $K_\$(0)$? Although we don't know what **S** will be worth at time T, we can determine this price. Consider the following two cases:

- If **K**'s current price were *higher* than the current stock price $S_\$(0)$, you could sell **K**, buy the stock **S** right now with only parts of the proceeds, hold it, and finally deliver it as promised. You could pocket the leftover money as immediate, risk-free gains.
- On the other hand, if the current price of **K** were *lower* than that of **S**, you could borrow the stock **S**, sell it at its current price, buy the contract **K** with only parts of the proceeds, and then wait unperturbed until **K** delivers you the **S** to be returned to its lender. Again, a profit at no risk.

The only price a buyer and seller can ever agree upon as fair is thus the current stock price:

$$K_\$(0) = S_\$(0) \quad \text{or} \quad K_S(0) = 1.$$

By the same reasoning, another contract $\mathbf{K}'$ delivering a bond **B** is priced as $K'_\$(0) = B_\$(0) = e^{-rT}$. Yet another contract $\mathbf{K}''$ delivering n units of **S** is of course priced as $K''_\$(0) = nS_\$(0)$.

These same simple relations do not hold for arbitrage assets like money. What would you be willing to pay now to get 1**\$** at time T? Certainly not 1**\$**. This is because money, as we mentioned, loses value with respect to other assets. Yet the workaround is simple, and we'll apply it in the following, second example.

With a *forward contract* **F**, you commit to buying **S** at time T for k**\$**:

$$\mathbf{F} \sim^T \mathbf{S} - k\$.$$

What would be a fair future exchange ratio or price k that obviates any upfront money exchange, i.e., that makes $F_\$(0) = 0$?

First, **F** is clearly simply the sum $\mathbf{K} + \mathbf{J}$, with $\mathbf{K} \sim^T \mathbf{S}$ and $\mathbf{J} \sim^T -k\$$. The current price of **K** is the same as the stock's (see our example immediately above):

$$K_\$(0) = S_\$(0).$$

But **J** is a monetary promise that ill-transcends time. Luckily, we know that a zero bond **B** can be exchanged for a **$** at maturity T:

$$\mathbf{B} \sim^T \mathbf{\$}.$$

This lets us express **J**'s promise in terms of a zero bond:

$$\mathbf{J} \sim^T -k\mathbf{\$} \sim^T -k\mathbf{B}.$$

We know that **J**'s current price must thus be $-k$ times the current price of the bond $B_\$(0)$. We get:

$$J_\$(0) = -kB_\$(0) \quad \text{or} \quad J_\$(0) = -ke^{-rT}.$$

For the current price of the forward **F** to be zero, we must have

$$0 = F_\$(0) = K_\$(0) + J_\$(0) = S_\$(0) - kB_\$(0) = S_\$(0) - ke^{-rT}.$$

This finally yields the so-called *forward price* k:

$$k = S_\$(0)e^{rT}.$$

(This forward price k is not to be confused with the price of the forward $F_\$(\cdot)$ itself. The latter is, by agreement upon k, zero at the beginning. As the stock price then starts to fluctuate, $F_\$(\cdot)$ will stray from zero and fluctuate as well.)

So we have settled both our initial questions by purely relying on price consistency or no-arbitrage. We now go one step further and explore the random nature of prices over time.

B.4 Prices as Expectations

Consider the stock price in terms of bonds, $S_B(\cdot)$. We know its current value $S_B(0)$, but future prices $S_B(T)$ are random and can only be described in terms of probabilities. The future price of a stock in terms of bonds can be higher or lower than the current price. However, as we hinted at before, it makes sense to assume the following: at least *on average*, $S_B(T)$ should not be higher or lower than $S_B(0)$. For if $S_B(T)$ were usually higher than the current $S_B(0)$, we would surely exchange all our bonds for stock, wait, and convert the stock back into bonds, because we would expect to often end up with more units of the bond than we set out with. In fact, everybody would try to enter such trades and thus drive up the current stock price. So we simply rule out such gains in our pricing model.

We treat $S_B(T)$ as a continuous random variable and assume it behaves according to the commonly used *log-normal* probability density, with x shorthand for $S_B(T)$:

$$p(x) = \frac{1}{x\sigma\sqrt{2T\pi}} e^{-\frac{(\log(x)-\log(S_B(0))+\frac{1}{2}\sigma^2 T)^2}{2\sigma^2 T}}.$$

As discussed before, we want the average of our random variable to be identical to the current stock price. Expressed via the expected value, we want $\mathbb{E}[S_B(T)]$ to be identical to $S_B(0)$. That this is indeed the case can be verified by computing the integral $\mathbb{E}[S_B(T)] = \int x\, p(x)\, dx$, which actually yields $S_B(0)$. So this seems to be a reasonable probability density.

We can generally view current prices as the expected value of asset units delivered. Take a contract that promises a random number X of bonds:

$$\mathbf{K} \sim^T X\mathbf{B}.$$

The price of $\mathbf{K}$ with respect to the delivered asset $\mathbf{B}$ must then reasonably be $\mathbb{E}[X]$ if we exclude arbitrage, otherwise we could gain or lose assets on average. A contract whose price we already derived may underline this point:

$$\mathbf{K} \sim^T \mathbf{S}.$$

At time T, we could immediately exchange the stock for bonds and consider this equivalent contract:

$$\mathbf{K} \sim^T S_B(T)\mathbf{B}.$$

This is a promise of a random number of bond units, and we therefore expect that

$$K_B(0) = \mathbb{E}[S_B(T)].$$

As noted above, this expected value evaluates to $S_B(0)$. The resulting dollar price of $\mathbf{K}$ thus coincides with our previous price derivation because

$$K_\$(0) = K_B(0)B_\$(0) = \mathbb{E}[S_B(T)]B_\$(0) = S_B(0)B_\$(0) = S_\$(0).$$

This is all fairly gimmicky when only considering trivial assets. Yet when pricing so-called *derivative* assets, whose promises are conditional on the prices of basic assets, we can gainfully apply the same approach. The most prominent such derivative is coming up.

B.5 The Call Option

We are now prepared to take on the call option **C**. It grants you the right to buy, at some future time T, a stock at a pre-determined strike price k:

$$\mathbf{C} \sim^T (\mathbf{S} - k\$)^+.$$

The $+$ denotes that you will exercise your claim and enter the buying transaction on the right only if the stock's dollar price is larger than the strike at time T, i.e., if the resulting portfolio value is positive.[12] The option expires worthlessly otherwise.

Just like with the forward, we first replace the **$** with a zero bond of maturity T:

$$\mathbf{C} \sim^T (\mathbf{S} - k\mathbf{B})^+.$$

We also replace the stock with a corresponding bond position:

$$\mathbf{C} \sim^T (S_B(T)\mathbf{B} - k\mathbf{B})^+.$$

As we are now only dealing with the bond asset on the right-hand side, we can factor it out:

$$\mathbf{C} \sim^T (S_B(T) - k)^+\mathbf{B}.$$

The coefficient of **B** is the random amount of units of **B** delivered by **C**, or $C_B(T)$. We are looking for the current price of the call, $C_B(0)$, which must equal

$$C_B(0) = \mathbb{E}[C_B(T)] = \mathbb{E}[(S_B(T) - k)^+].$$

We next have to actually calculate the corresponding integral, with $x = S_B(T)$:

$$\begin{aligned} C_B(0) &= E[(x-k)^+] \\ &= \int_{-\infty}^{\infty} (x-k)^+ p(x)\,dx \\ &= \int_{k}^{\infty} (x-k) p(x)\,dx \\ &= \int_{k}^{\infty} (x-k) \frac{1}{x\sigma\sqrt{2T\pi}} e^{-\frac{(\log(x)-\log(S_B(0))+\frac{1}{2}\sigma^2 T)^2}{2\sigma^2 T}}\,dx. \end{aligned}$$

[12] The $(\cdot)^+$ is a valid mathematical operator on asset expressions because the *sign* of a portfolio's price does not depend on the reference asset used for pricing.

Depending on your mood, you can integrate this expression by hand or use integration software like Mathematica. The integral evaluates to:

$$C_B(0) = S_B(0)\ \Phi\left[\frac{1}{\sigma\sqrt{T}}(\log S_B(0) - \log K + \frac{1}{2}\sigma^2 T)\right]$$
$$-K\ \Phi\left[\frac{1}{\sigma\sqrt{T}}(\log S_B(0) - \log K - \frac{1}{2}\sigma^2 T)\right].$$

This almost looks like the formula you find in the books. To exactly match that classic formulation, which is given in dollar and not bond terms, two additional steps are required. First, we replace the stock price in bond terms with the equivalent dollar expression. We have

$$S_\$(0) = S_B(0)B_\$(0) = S_B(0)e^{-rT} \implies S_B(0) = S_\$(0)e^{rT},$$

and of course

$$\log S_B(0) = \log(S_\$(0)e^{rT}) = \log S_\$(0) + rT.$$

If we also translate the call price from bond to dollar terms via

$$C_\$(0) = C_B(0)B_\$(0) = C_B(0)e^{-rT},$$

we obtain the classic Black-Scholes formula:

$$C_\$(0) = S_\$(0)\ \Phi\left[\frac{1}{\sigma\sqrt{T}}(\log S_\$(0) - \log K + rT + \frac{1}{2}\sigma^2 T)\right]$$
$$-Ke^{-rT}\ \Phi\left[\frac{1}{\sigma\sqrt{T}}(\log S_\$(0) - \log K + rT - \frac{1}{2}\sigma^2 T)\right].$$

As much fun as this is, such formulas are rarely used for pricing. Options are traded, and their prices are determined by supply and demand. We can consider them a given like stock or bond prices. The main use we have for this framework is that we can, if you will, reverse it and determine the value of σ that yields the known option price—this σ is called the *implied volatility*. Just like interest rates in the context of bonds, it serves as a convenient way of quoting option prices.

B.6 Views on Probabilities

This is of course just a very brief glimpse into pricing. One additional facet worth hinting at, though, are the probabilities involved. To illustrate their behavior, we look at a simplified pricing model where the prices of a stock and a bond evolve into

only two states (u for "stock up" and d for "stock down") after some time T:

$$S_\$(0) = 10 \begin{cases} \nearrow S_\$(T) = 20 \\ \searrow S_\$(T) = 5 \end{cases} \qquad B_\$(0) = 0.9 \begin{cases} \nearrow B_\$(T) = 1 \\ \searrow B_\$(T) = 1 \end{cases}$$

We are mainly interested in prices with respect to no-arbitrage assets. Here are all asset prices with respect to the bond:

$$S_B(0) = 11.11 \begin{cases} \nearrow S_B(T) = 20 \\ \searrow S_B(T) = 5 \end{cases} \qquad B_B(0) = 1 \begin{cases} \nearrow B_B(T) = 1 \\ \searrow B_B(T) = 1 \end{cases}$$

This is the view we adopted in pricing the option above. But of course we can also express the asset prices with respect to the stock—unlike us, this is how Bill Gates might view the world:

$$S_S(0) = 1 \begin{cases} \nearrow S_S(T) = 1 \\ \searrow S_S(T) = 1 \end{cases} \qquad B_S(0) = 0.09 \begin{cases} \nearrow B_S(T) = 0.05 \\ \searrow B_S(T) = 0.20 \end{cases}$$

What probabilities p_u and $p_d = 1 - p_u$ should we—in our bond view—assign to the two outcomes? Ruling out arbitrage tells us:

$$S_B(0) = 11.11 = \mathbb{E}[S_B(T)] = p_u \times 20 + (1 - p_u) \times 5 \implies p_u = 0.407.$$

How about Bill Gates? He wants to assume the following:

$$B_S(0) = 0.09 = \mathbb{E}[B_S(T)] = p_u \times 0.05 + (1 - p_u)0.20 \implies p_u = 0.733.$$

Whoa—the probabilities differ! We see that depending on the reference asset used, the no-arbitrage condition entails different probabilities. We'd best rename those distinct probabilities for the "stock up" scenario to p_u^B for our bond-based view and to p_u^S for Bill's stock-based one. We end up with two ways of computing the expectations involved:

$$\mathbb{E}^B[X] = p_u^B x_u + (1 - p_u^B)x_d,$$
$$\mathbb{E}^S[X] = p_u^S x_u + (1 - p_u^S)x_d.$$

We have, by construction,

$$\mathbb{E}^B[S_B(T)] = 11.11 = S_B(0),$$
$$\mathbb{E}^S[B_S(T)] = 0.09 = B_S(0),$$

as well as

$$\mathbb{E}^B[B_S(T)] = 0.14 \neq B_S(0),$$
$$\mathbb{E}^S[S_B(T)] = 16 \neq S_B(0).$$

How to price a contract **C** that pays out the stock **S** in the "stock up" scenario and nothing in the "stock down" one? Under the bond view, getting **S** is identical to getting $S_B(T)\mathbf{B}$, and the payoff (in bond terms) of this contract at time T is thus

$$C_B(T) = \begin{cases} S_B(T) = 20 & \text{in the "stock up" scenario,} \\ 0 & \text{otherwise.} \end{cases}$$

Its current price is

$$C_B(0) = \mathbb{E}^B[C_B(T)] = p_u^B \times 20 + (1 - p_u^B) \times 0 = 8.15.$$

Under the stock view, the contract payoff at time T is even simpler:

$$C_S(T) = \begin{cases} 1 & \text{in the "stock up" scenario,} \\ 0 & \text{otherwise.} \end{cases}$$

We have

$$C_S(0) = \mathbb{E}^S[C_S(T)] = p_u^S \times 1 + (1 - p_u^S) \times 0 = 0.73.$$

Yet both views agree on the dollar price:

$$C_\$(0) = C_B(0)B_\$(0) = 8.15 \times 0.9 = 7.33,$$
$$C_\$(0) = C_S(0)S_\$(0) = 0.73 \times 10 = 7.33.$$

The more natural way to price such a stock-affine payoff is Bill's stock view. Although multiplying by 20 in the bond view is certainly doable here, this step falls away for Bill. Hopefully, this lets you imagine that in the continuous case, where we have to evaluate integrals, a suitable problem formulation can bring about considerable simplifications.

We can use much the same reasoning with our call option. We have used a (bond-based) probability density that made sure that $S_B(0) = \mathbb{E}^B[S_B(T)]$. (Note: we usefully renamed the expectation just like above.) There is an alternative density for $S_B(\cdot)$ that allows us to evaluate expectations under the stock view as well, i.e., expressions of the form $\mathbb{E}^S[f(S_B(T))]$.[13]

[13] There are also two densities for $B_S(\cdot)$, corresponding to the two expectations. One of them neatly makes sure that $B_S(0) = \mathbb{E}^S[B_S(T)]$.

Expressing the call payoff via the indicator function as in

$$\mathbf{C} \sim^{T} \mathbb{1}_{S_B(T)\geq k}\mathbf{S} - k\mathbb{1}_{S_B(T)\geq k}\mathbf{B}$$

lets us then derive the call's current price from

$$\mathbf{C} \sim^{0} \mathbb{E}^{S}[\mathbb{1}_{S_B(T)\geq k}]\mathbf{S} - k\,\mathbb{E}^{B}[\mathbb{1}_{S_B(T)\geq k}]\mathbf{B}.$$

Possibly even niftier: we can also selectively use $B_S(T)$ in this expression (recall that $B_S = 1/S_B$ and that the reciprocal of a log-normal distribution is also log-normal and helpfully preserves σ) and thereby use the "canonical" distributions under each expectation:

$$\mathbf{C} \sim^{0} \mathbb{E}^{S}[\mathbb{1}_{B_S(T)\leq \frac{1}{k}}]\mathbf{S} - k\,\mathbb{E}^{B}[\mathbb{1}_{S_B(T)\geq k}]\mathbf{B}.$$

Computing these expectations also yields the Black-Scholes formula.

Further Reading

C

An excellent book about the basis of it all—debt and money—is Graeber's "Debt: The First 5000 years" (Graeber 2014), which outlines how debt preceded and indeed paved the way for money and the subsequent financial products and markets. Some insight into why those markets may behave the way they do can be found in Akerlof and Shiller's "Animal Spirits" (Akerlof and Shiller 2010).

An extensive market risk classic is Jorion's "Value at Risk" (Jorion 2006), and many other general reference resources are available online.[14] An overview of the risk landscape and the particular role of market risk in it is given in Allen (2009). Many of the core concepts compiled in the book you are holding can be found in Ortega et al. (2009), a paper by my former work colleagues and creators of the initial version of our scenario generator. The historical VaR approach championed in this book belongs to the family of filtered historical simulations (Barone-Adesi et al. 1999, 2008). The BRW model is a commonly-encountered alternative (Boudoukh et al. 1998).

Artzner et al. (1999) shine some light on desirable properties of risk measures and introduce the influential concept of *coherent* measures. An in-depth treatment of risk measures' verificability can be found in Ziegel (2014). A workaround for the usually unstable additive decomposition of VaR to individual positions is presented in Epperlein and Smillie (2006). Anyone using p-values to make a point might find (Wasserstein and Lazar 2016) useful.

In the context of a VaR model, you'll inevitably encounter issues of pricing and arbitrage, topics we hinted at only very briefly. A great gateway into this world is Jan Vecer's "Stochastic Finance: A Numeraire Approach" (Vecer 2011). He neatly distinguishes between assets and their prices, concepts often intermingled in traditional notations. He also doesn't dwell on technical details and emphasizes explicit step-by-step calculations. Then either head down the math alley with Shreve's excellent books, especially (Shreve 2008), or get a comprehensive and less

[14] www.value-at-risk.net.

M. Auer, *Hands-On Value-at-Risk and Expected Shortfall*, Management for Professionals, https://doi.org/10.1007/978-3-319-72320-4

formal overview on pricing with Hull's standard reference "Options, Futures and Other Derivatives" (Hull 2011).

Books by practitioners can then greatly help you with more arcane products (Zhang 1996), tricky issues of calibration to market data (Rebonato 2002), and explicit algorithms (Brigo and Mercurio 2007). Supplement your modeling skills with the invaluable (Kutner et al. 2004). Finally, make sure to check out Glasserman's superb "Monte Carlo Methods in Financial Engineering" (Glasserman 2003). It is very accessible, and many of the presented methods, e.g., variance reduction techniques, can not only be used in pricing but also in our simple VaR model setup.

If you want to expose yourself to the wide and fast-paced IT-field, it can't hurt to understand its slang. Browse, for example, through the table of contents in Sommerville (2015), and try to zoom in on unfamiliar terms until your have a grasp of their meaning. Soon you should be able to roughly decipher the programmers' gobbledygook ("we have deployed unit testing to the grid"). For managing IT projects, consider looking into *agile software development* (Martin 2002).

Then learn about the Linux operating system (you can install one on a virtual machine[15] on your Windows desktop) and familiarize yourself with its command line interface (Powers et al. 2002). To actually learn how to program, start off with the programming language C, best with the concise and very elegant (Kernighan and Ritchie 1989). Once you master the concept of pointers, feel free to speed up your progress by learning Python (Gaddis 2014), possibly via some of the excellent online courses available.[16] Python also allows you to learn about object-oriented programming. Once you understand why a "square" class should not inherit from the "rectangle" one, you are ready for C++ (Stroustrup 2013), design patterns (Gamma et al. 1994), and UML (Fowler 2004). A tool for creating UML diagrams—high-level representations of object-oriented code—is UMLet.[17]

As for mathematical and statistical support tools, definitely check out NumPy[18] (a Python add-on) or R.[19] (NumPy, unlike R, uses 0-based indexing, which is better.[20]) Many of the examples in this book can be reenacted in Excel or via supporting Monte Carlo add-ins like MonteCarlito.[21] Finally, drop by at this book's www.value-at-risk.com.

[15] www.virtualbox.org.

[16] www.codecademy.com/learn/python.

[17] www.umlet.com (full disclosure: tool by author).

[18] www.numpy.org.

[19] www.r-project.org.

[20] www.cs.utexas.edu/users/EWD/transcriptions/EWD08xx/EWD831.html—or google "Edsger Dijkstra why numbering should start at zero" should this link prove unstable.

[21] www.montecarlito.com (tool by author).

References

Akerlof, G. A., & Shiller, R. J. (2010). *Animal spirits*. Princeton University Press: Princeton.

Allen, S. L. (2009). *Financial risk management*. Wiley: Hoboken.

Artzner, P., Delbaen, F., Eber, J. -M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*(3) , 203–228.

Barone-Adesi, G., Engle, R., & Mancini, L. (2008). A GARCH option pricing model with filtered historical simulation. *Review of Financial Studies*, *21*(3), 1223–1258.

Barone-Adesi, G., Giannopoulos, K., & Vosper, L. (1999). VaR without correlations for non-linear portfolios. *Futures Markets*, *19*, 583–602.

Boudoukh, J., Richardson, M., & Whitelaw, R. (1998). The best of both worlds. *Risk*, *11*, 64–67.

Brigo, D., & Mercurio, F. (2007). *Interest rate models*. Springer: Berlin

Epperlein, E., & Smillie, A. (2006). Cracking VAR with kernels. *Risk*, *19*(8), 70–74.

Fowler, M. (2004). *UML distilled: A brief guide to the standard object modeling language*. Addison-Wesley: Boston.

Gaddis, T. (2014). *Starting out with python*. Pearson: Boston.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994) *Design patterns: Elements of reusable object-oriented software*. Addison-Wesley: Boston.

Glasserman, P. (2003) *Monte Carlo methods in financial engineering*. Springer: New York.

Graeber, D. (2014). *Debt: The first 5000 years*. Melville House: Brooklyn.

Hull, J. C. (2011) *Options, futures and other derivatives*. Prentice Hall: Upper Saddle River.

Jorion, P. (2006) *Value at risk*. McGraw-Hill: New York.

Kernighan, B. W., & Ritchie, D. (1989) *The C programming language*. Prentice Hall: Upper Saddle River.

Kutner, M., Nachtsheim, C., & Neter, J. (2004) *Applied linear regression models*. McGraw-Hill: New York.

Martin, R. C. (2002) *Agile software development: Principles, patterns, and practices*. Pearson: Upper Saddle River.

Ortega, J. -P., Pullirsch, R., Teichmann, J., & Wergieluk, J. (2009). A new approach for scenario generation in risk management. preprint arXiv:0904.0624.

Powers, S., Peek, J., O'Reily, T., & Loukides, M. (2002). *Unix power tools*. O'Reilly: Sebastopol.

Rebonato, R. (2002). *Modern pricing of interest-rate derivatives*. Princeton University Press: Princeton.

Shreve, S. E. (2008). *Stochastic calculus for finance II: Continuous-time models*. Springer: New York.

Sommerville, I. (2015). *Software engineering*. Pearson: Upper Saddle River.

Stroustrup, B. (2013). *The C++ programming language*. Addison-Wesley: Boston.

Vecer, J. (2011). *Stochastic finance: A numeraire approach*. CRC Press: Boca Raton.

M. Auer, *Hands-On Value-at-Risk and Expected Shortfall*, Management for Professionals, https://doi.org/10.1007/978-3-319-72320-4

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician* (preprint online).

Zhang, P. G. (1996). *Exotic options: A guide to second generation options*. World Scientific Publishing: Singapore.

Ziegel, J. F. (2014). Coherence and elicitability. preprint arXiv:1303.1690.

Index

M. Auer, *Hands-On Value-at-Risk and Expected Shortfall*, Management for Professionals, https://doi.org/10.1007/978-3-319-72320-4