Math Tools: Random Variables

Revised: October 4, 2014

Our first set of math tools is concerned with randomness: how we think about it, how we use it, and so on. It's an essential idea in macroeconomics and finance — in fact in the world at large — and the tools developed to work with it are incredibly useful. Remind yourself as we work through this material: Tools make life easier, even if it takes some time to learn how to use them. The plan is to describe some concepts and tools and put them to work. We won't worry too much about mathematical niceties, but you can find more formal versions elsewhere.

We'll use three concepts to describe randomness: random variables, distributions, and moments. I know this sounds a little circular, but *random variables* describe things that are random. They are characterized by their *distributions*, the probabilities of possible outcomes. *Moments* are summary measures that tell us something about the distribution.

Moments are, in some ways, crude measures of distributions, but we use them to describe two sets of properties. The mean and variance reflect a random variable's location (where it is on the real line) and dispersion (how spread out it is). If the world is normal, then that's all we need. But of course lots of things in this world aren't normal. Skewness and kurtosis reflect the shape of the distribution and helpful to identify deviations from the normal distribution. They will show up when we look at risk premiums, option prices, and other things.

1 Random variables and probability distributions

Random variables are used to describe things that are random, typically in the sense that we don't know their outcomes ahead of time. You might think of the weather (rain or shine?), the economy (boom, bust, or somewhere in between?), the stock market (ditto), or sports (will the Steelers win?).

To make this precise, it's helpful to build the definition from components. This will start off abstract, but turn concrete quickly — and stay that way. The components are:

- *States.* Let's start with what we'll call a *state*, one of the possible outcomes of a random process. You may see the terms *outcome* or *event* used instead, but state serves our purposes. We represent a state with the letter z and the set of all possible states — the *state space* — by $\mathcal{Z}$. [Draw event tree here.] Sometimes state $z = 1$ occurs, sometimes state $z = 2$, and so on. Part of the art of applied work is to come up with a useful practical definition of $\mathcal{Z}$. If we're talking about the weather, the states might be $z =$ rain and $z =$ shine. If we're talking about the stock market, we might assign different states to every possible value of the S&P 500 on (say) the last day of the year.

- *Probabilities.* For each state z, we assign a number $\text{Prob}(z) = p(z)$ that we refer to as the *probability* of z. Here $\text{Prob}(A)$ means (literally) the probability of some arbitrary subset A of $\mathcal{Z}$ and p is a function of the state z. We call the complete collection of $p(z)$'s the *probability distribution*. Not every set of numbers works. Legitimate probabilities must be nonnegative and sum to one.

- *Random variables.* A *random variable* is a function that assigns a real number to every state: $x(z)$. Note that x inherits randomness from z so it is, precisely, a random variable. Sometimes people distinguish between the random variable and the values the random variable takes, but we'll use x for both.

In practice we often ignore z and define probabilities directly over x, but we'll see later on that there are times when the distinction between states and random variables is helpful.

Some common examples of probability distributions of random variables:

- *Bernoulli.* The state space $\mathcal{Z}$ has two elements: $\mathcal{Z} = \{z_1, z_2\}$. If we're flipping a coin, z_1 might represent heads and z_2 tails. A random variable assigns numbers to these two states. The simplest version uses zero and one: $x(z_1) = 0$, $x(z_2) = 1$. The probabilities $p(z_1)$ and $p(z_2)$ are nonnegative numbers that sum to one.

- *Poisson.* This is a little more complicated, but it's a beautiful example that's often used in finance (many other places, too, of course). Suppose $\mathcal{Z}$ consists of the nonnegative integers: $\{0, 1, 2, \ldots\}$. The probability of any particular z is

$$p(z) \quad = \quad e^{-\omega}\omega^z/z!,$$

  with parameter $\omega > 0$ ("intensity").

  Are they legitimate probabilities? Well, they're all positive, so we're good there. Do they sum to one? That's more complicated. The exponential function has the power series expansion

$$e^x \quad = \quad 1 + x + x^2/2 + x^3/3! + x^4/4! + \cdots \quad = \quad \sum_{j=0}^{\infty} x^j/j!. \tag{1}$$

  (This is the Taylor series representation of e^x at the point $x = 0$.) Our probabilities have a similar form:

$$\sum_{z=0}^{\infty} p(z) \quad = \quad e^{-\omega} \sum_{z=0}^{\infty} \omega^z/z! \quad = \quad e^{-\omega}e^{\omega} \quad = \quad 1.$$

  So they are, in fact, legitimate probabilities.

- *Normal (Gaussian).* Here we'll let the state space be the real line. We'll also set $x = z$, which allows us to ignore z. We refer to such random variables as continuous, to distinguish them from random variables that based on a discrete set of states, such as the two we just looked at.

For continuous random variables, we describe probabilities with what's called a *probability density function* or pdf $p(x)$. Probabilities over an interval $[a, b]$ are integrals,

$$\text{Prob}(a \leq x \leq b) \quad = \quad \int_a^b p(x)dx.$$

The function $p(x)$ must be nonnegative for all values of x and integrate to one,

$$\int_{-\infty}^{\infty} p(x)dx \quad = \quad 1,$$

the analog of the sum we used earlier. The counterpart of $p(x)$ in the discrete case is $p(x)dx$ — it's important to include the dx.

A *normal* or Gaussian random variable has density function

$$p(x) \quad = \quad (2\pi\sigma^2)^{-1/2} \exp[-(x-\mu)^2/(2\sigma^2)],$$

the well-known "bell-shaped curve." [If you graph this, you'll see why.] It's positive for all x and integrates to one, although (for good reason) we'll take the latter on faith rather than demonstrate it. The two parameters are μ and σ^2 so we often write $x \sim \mathcal{N}(\mu, \sigma^2)$ to mean "x is normal with parameters μ and σ^2." We'll see shortly what these parameters do. The so-called *standard normal* refers to the case $\mu = 0$ and $\sigma = 1$. [Graph $p(x)$ and show how it changes when we change μ and σ.]

There are lots of other common distributions of random variables, many of which are summarized in Wikipedia: search "normal distribution" or "Bernoulli distribution." (I know Wikipedia has a bad rep, but the math content is pretty good.)

2 Expectations and moments

The behavior of a random variable is described completely by (i) the set of values it can take and (ii) their probabilities. Sometimes that's too much information: we'd like a smaller number of properties that capture some salient features. I like looking at the whole distribution, or even better a picture of it, but summary numbers can be useful, too.

We start with the concept of an *expectation*. If x is a random variable with probabilities p, its expectation is the weighted average value of x using probabilities as weights:

$$E(x) \quad = \quad \sum_z x(z)p[x(z)].$$

For a continuous random variable, we replace the sum with an integral. The notation is a little cumbersome, which is why we just write $E(x)$ most of the time.

We can extend this concept to any function of x:

$$E[f(x)] \quad = \quad \sum_z f[x(z)]p[x(z)]. \tag{2}$$

There may be cases where the sum (or integral) doesn't converge, but we won't worry about that now — or ever, really.

We see from the definition (2) that expectations are linear, which means they satisfy these two properties. (i) If f is a sum, so that $f(x) = g_1(x) + g_1(x)$ for two functions g_1 and g_2, then its expectation is the sum of the expectations of its components. In short-hand notation,

$$E[f(x)] \quad = \quad E[g_1(x)] + E[g_2(x)].$$

(ii) If we multiply f by a constant c, then its expectation is c times the expectation of the original function:

$$E[cf(x)] \quad = \quad cE[f(x)].$$

Together they imply that the expectation of a linear function of x is the same linear function of its expectation:

$$E(a + bx) \quad = \quad a + bE(x).$$

We'll use these properties so often they'll become second nature.

Here's a well-known application. A *moment* is the expectation of a power of x:

$$\mu'_j \quad = \quad E(x^j)$$

for j a positive integer. The first one (μ'_1) is known as the *mean*. It's a measure of the "location" of the probability distribution. If we change the mean, the probability distribution shifts left or right. Think about graphing the probability distribution of $x + a$ for different values of a. If the mean of x is μ'_1, then the mean of $x + a$ is $\mu'_1 + a$, so when we move the distribution back and forth by changing a, that's reflected in the mean.

We often use *central moments* instead, meaning we look at powers of x minus its mean:

$$\mu_j \quad = \quad E[(x - \mu'_1)^j]$$

The idea is to take location out of the calculation. The first central moment is zero by construction:

$$\mu_1 \quad = \quad E(x - \mu'_1) \quad = \quad E(x) - \mu'_1 \quad = \quad \mu'_1 - \mu'_1 \quad = \quad 0.$$

(Here and elsewhere: If any of these steps seem mysterious, write out the calculation of the expectation for the discrete case.)

The second central moment is called the *variance*. It's a measure of dispersion: how spread out or dispersed the distribution is. We denote it by

$$\mathrm{Var}(x) \quad = \quad E[(x - \mu'_1)^2].$$

Since squares are positive (or at least nonnegative), so is the variance. If we write out the definition of the variance, we see it can be expressed in terms of (noncentral or raw) moments:

$$E[(x-\mu_1')^2] \;=\; E[x^2 - 2x\mu_1' + (\mu_1')^2] \;=\; \mu_2' - 2(\mu_1')^2 + (\mu_1')^2 \;=\; \mu_2' - (\mu_1')^2.$$

If we wanted to, we could compute the variance this way. That's pretty common, but we'll see shortly there's a better (by which I mean easier) way. The *standard deviation* is the (positive) square root of the variance and is often used the same way.

Let's see how the mean and standard deviation reflect location and scale, respectively. Suppose x has given mean and standard deviation. What are the mean and standard deviation of $y = a + bx$? The mean is $E(y) = a + bE(x)$, so a moves it up and down. The variance is

$$\begin{aligned}\mathrm{Var}(x) &= E\left\{[y - E(y)]^2\right\} \;=\; E\left\{[a + bx - a - bE(x)]^2\right\} \\ &= E\left\{[bx - bE(x)]^2\right\} \;=\; b^2 E\left\{[x - E(x)]^2\right\} \;=\; b^2\mathrm{Var}(x).\end{aligned}$$

The standard deviation is the positive square root. If σ is the standard deviation of x, then $|b|\sigma$ is the standard deviation of y. (Why absolute value? The standard deviation is the positive square root of the variance.) Note that all of this follows from applying definitions.

Let's go back to our examples and see what the mean and variance are:

- *Bernoulli.* Let the probability that $x = 1$ be ω and the probability that $x = 0$ be $1 - \omega$. How do we find the mean and variance? The easiest way is to look them up in Wikipedia (search "Bernoulli distribution"), but let's see if we can find them on our own. The mean is (apply the definition)

  $$E(x) \;=\; (1-\omega)\cdot 0 + \omega \cdot 1 \;=\; \omega.$$

  We find the variance from the second moment:

  $$E(x^2) \;=\; (1-\omega)\cdot 0^2 + \omega \cdot 1^2 \;=\; \omega.$$

  The variance is therefore

  $$\mathrm{Var}(x) \;=\; E(x^2) - [E(x)]^2 \;=\; \omega - \omega^2 \;=\; \omega(1-\omega).$$

  The standard deviation is the (positive) square root of this.

- *Poisson.* The mean is

  $$E(x) \;=\; e^{-\omega}\sum_{j=0}^{\infty} j\omega^j/j! \;=\; e^{-\omega}\omega\sum_{j=1}^{\infty}(j-1)\omega^{j-1}/(j-1)! \;=\; \omega.$$

  [You'll have to think about this a little – or better yet, wait a few minutes and we'll derive this by an easier route.] The second moment is

  $$E(x^2) \;=\; e^{-\omega}\sum_{j=0}^{\infty} j^2\omega^j/j! \;=\; ??.$$

  We could fight our way through this, but since an easier way is just around the corner, I'll surrender now and come back to fight another day.

- *Normal.* We'll postpone this one, too.

Our last topic here is *sample moments*: moments computed from data. The idea is to use sample weights rather than probabilities. Given a sample of x_t's for $t = 1, 2, \ldots, T$, the sample mean is

$$\bar{x} \quad = \quad T^{-1} \sum_{t=1}^{T} x_t.$$

Similarly, the jth sample moment is

$$T^{-1} \sum_{t=1}^{T} x_t^j.$$

The jth sample *central* moment is

$$T^{-1} \sum_{t=1}^{T} (x_t - \bar{x})^j.$$

If $j = 2$ we get the sample variance. And if we take the square root, we get the sample standard deviation. (And yes, we divide by T, not $T - 1$ or something else.)

We use sample moments the same way we use moments: to describe the distribution. Summary statistics include the mean, the variance, the standard deviation, and so on. If the x_t's are produced by a specific distribution, then with enough data we would hope that the sample moments will be "close" to the moments of the distribution that generated them.

3 Generating functions

Next up is one of my favorite tools: generating functions. It's a tool with a wide range of uses, but we're interested in one: as a short-cut in computing moments. If you've run across Laplace, Fourier, or z transforms, they're closely related.

The *moment generating function* (mgf) is defined by

$$h(s) \quad = \quad E\left(e^{sx}\right), \tag{3}$$

a function of the real number s. (It's common to use t instead of s, but we need t for time.) Note that $h(0) = 1$. [Ask yourself why.]

The mgf is a tool, like a hammer, and we hammer things with it. If we hammer probability distributions, we get moments as a byproduct. Recall the Taylor series expansion (1) of the exponential function. If we expand e^{sx} the same way and take expectations, the moments pop out:

$$\begin{aligned} h(s) &= E\left[1 + (sx) + (sx)^2/2 + (sx)^3/3! + \cdots\right] \\ &= 1 + \mu_1' s + \mu_2'(s^2/2) + \mu_3'(s^3/3!) + \cdots. \end{aligned}$$

With a little more insight, we see that we can recover the moments by differentiating h and setting $s = 0$. The first derivative is the mean:

$$h^{(1)}(0) \;=\; \left.\frac{dh(s)}{ds}\right|_{s=0} \;=\; \mu_1'.$$

Here $h^{(1)}(0)$ means the first derivative of the function $h(s)$ evaluated at $s = 0$. Similarly, high-order moments follow from high-order derivatives:

$$h^{(j)}(s) \;=\; \left.\frac{d^j h(s)}{ds^j}\right|_{s=0} \;=\; \mu_j'$$

This looks horrible, but it just says that the jth moment is the jth derivative evaluated at $s = 0$. Bottom line: if we know the mgf, we can find moments by differentiating it. Better yet, we can get Matlab to do the differentiating.

Let's go back to our examples:

- *Bernoulli.* The mgf is

  $$h(s) \;=\; (1-\omega)e^{s\cdot 0} + \omega e^{s\cdot 1} \;=\; (1-\omega) + \omega e^s.$$

  The first two derivatives give us the first two (noncentral) moments:

  $$\begin{aligned} h^{(1)}(0) &= \omega = \mu_1' \\ h^{(2)}(0) &= \omega = \mu_2'. \end{aligned}$$

  The variance is therefore $\mu_2' - (\mu_1')^2 = \omega(1-\omega)$, as we saw earlier.

- *Poisson.* The mgf is

  $$h(s) \;=\; \sum_{z=0}^{\infty} e^{sz} e^{-\omega}\omega^z/z! \;=\; e^{-\omega}\sum_{z=0}^{\infty} \left(e^s\omega\right)^z/z! \;=\; e^{-\omega}e^{e^s\omega} \;=\; e^{\omega(e^s-1)}.$$

  The first two derivatives are

  $$\begin{aligned} h^{(1)}(0) &= \omega \\ h^{(2)}(0) &= \omega + \omega^2. \end{aligned}$$

  The mean and variance are therefore both equal to ω.

- *Normal.* We find the mgf by completing the square. Pay attention here: we'll see this over and over again, including our derivation of the Black-Scholes-Merton option pricing formula. We start with the definition:

  $$h(s) \;=\; (2\pi\sigma^2)^{-1/2}\int_{-\infty}^{\infty} e^{sx}e^{-(x-\mu)^2/2\sigma^2}dx.$$

  The exponents are

$$\begin{aligned} sx - (x-\mu)^2/2\sigma^2 &= -(1/2\sigma^2)\left[-2\sigma^2 sx + x^2 - 2\mu x + \mu^2\right] \\ &= \mu s + \sigma^2 s^2/2 - [x - (\mu + s\sigma^2)]^2/2\sigma^2. \end{aligned} \tag{4}$$

This may take you a couple minutes, but try expanding both expressions and lining up terms. I put a box around it, because it's an equation we'll see again, specifically when we come to option prices.

When we plug the result into the integral and rearrange terms, we have

$$h(s) = e^{\mu s + \sigma^2 s^2/2} (2\pi\sigma^2)^{-1/2} \int_{-\infty}^{\infty} e^{-[x-(\mu+s\sigma^2)]^2/2\sigma^2} dx.$$

The last term is a normal density function and therefore integrates to one. [We didn't prove this, but we know pdf's must integrate to one. Right?] That leaves us with

$$h(s) = e^{\mu s + \sigma^2 s^2/2}.$$

If you differentiate, you can show that this implies a mean of μ and a variance of σ^2.

The moment generating function gives us, in these and many other cases, an easy route to finding moments. The *cumulant generating function* (cgf) is even better. It's defined as the logarithm of the moment generating function:

$$k(s) = \log h(s). \tag{5}$$

Note well: In this class, and in Matlab, log means the natural or base-e logarithm. Always.

The cgf has the Taylor series expansion

$$k(s) = \kappa_1 s + \kappa_2 s^2/2 + \kappa_3 s^3/3! + \cdots, \tag{6}$$

where κ_j is the jth derivative of $k(s)$ at $s = 0$:

$$\kappa_j = k^{(j)}(0).$$

[Question: Why is there no $k(0)$ term?] We refer to these derivatives as *cumulants* κ_j.

The question is what the cumulants are, other than derivatives of the cgf. With the moment generating function, we could see that the coefficients were raw moments. But what are they here? The best way to answer that is to connect the cumulants to moments, which we do by linking derivatives of k to those of h using (5). For example, the first two derivatives are

$$\begin{aligned} k^{(1)}(0) &= h^{(1)}(0)/h(0) = h^{(1)}(0) \\ k^{(2)}(0) &= h^{(2)}(0) - h^{(1)}(0)^2, \end{aligned}$$

the mean and variance.

Note that the second cumulant "centralizes" the second moment: we get the variance directly, rather than the second noncentral moment. That's an example of a more general property. The cgf of $y = a + bx$ is

$$k(s; y) \quad = \quad as + k(sb; x).$$

[This may look mysterious, but just apply the definition: $h(s; y) = E[\exp(as + bsx)] = \exp(as)E[\exp(bsx)] = \exp(as)h(bs; y)$ and take logs.] As we've seen, a changes the location and b changes the scale. The cumulants of y are therefore connected to those of x by

$$\begin{aligned} \kappa_1(y) &= a + b\kappa_1(x) \\ \kappa_j(y) &= b^j \kappa_j(x) \quad \text{for } j = 2, 3, \ldots \end{aligned}$$

That is: after the first one, cumulants aren't affected by location, and scale shows up as a power.

After the mean and variance, the most useful moments/cumulants are the third and fourth measuring, respectively, *skewness* and *kurtosis*. Skewness refers to the asymmetry of the distribution: odd cumulants (and central moments) are zero after the first for any symmetric distribution. [Can you show this?] Kurtosis (sort of) refers to how much weight is in the tails of the density; holding the mean and variance constant, a distribution with greater kurtosis will have more weight in the tails and, to keep the variance constant, more near the center as well. There's no theorem to that effect, but it's a useful statement nonetheless. [Draw a picture.]

The standard measures of skewness and kurtosis are based on the third and fourth cumulants:

$$\begin{aligned} \gamma_1 &= \kappa_3/(\kappa_2)^{3/2} \quad \text{(skewness)} \\ \gamma_2 &= \kappa_4/(\kappa_2)^2 \quad \text{(excess kurtosis)} \end{aligned}$$

The denominators take care of scaling. [If you don't see this, note how scaling affects cumulants, and therefore γ_1 and γ_2.] We'll explain the term "excess" shortly.

Our examples again:

- *Bernoulli.* The cgf is

  $$k(s) \quad = \quad \log\left[(1 - \omega) + \omega e^s\right].$$

  For practice: compute the first four cumulants and the measures of skewness and excess kurtosis.

- *Poisson.* The cgf is

  $$k(s) \quad = \quad \omega(e^s - 1).$$

  Its derivatives are all the same, so we have

  $$\kappa_j \quad = \quad \omega$$

  for all $j \geq 1$. Skewness is $\gamma_1 = \omega/\omega^{3/2} = \omega^{-1/2} > 0$. Excess kurtosis is $\gamma_1 = \omega/\omega^2 = \omega^{-1} > 0$.

- *Normal.* The cgf is

$$k(s) = \mu s + \sigma^2 s^2/2.$$

What's wonderful about this is that all cumulants after the first two are zero. Any nonzero cumulants beyond that are signs that the distribution isn't normal. Skewness and excess kurtosis are examples of that.

One last thing: Why do we say "excess" kurtosis? An alternative measure of kurtosis follows from the fourth central moment:

$$\mu_4/(\mu_2)^2.$$

Since $\mu_2 = \kappa_2$, only the numerator differs from our earlier measure. And since neither measure depends on location or scale, it's enough to look at the standard normal, which has $\mu = 0$ and $\sigma = 1$. The mgf is therefore $h(s) = e^{s^2/2}$ and its fourth derivative is

$$h^{(4)}(s) = 3e^{s^2/2} + 6s^2 e^{s^2/2} + s^4 e^{s^2/2}.$$

[I did this in Matlab because I'm lazy.] The fourth moment (central because the mean is zero) is therefore $h^{(4)}(0) = 3$. Why 3? That's just the way it is. But it tells us that if the kurtosis of a normal random variable is three, we need to subtract three to detect departures from normality. That's what γ_2 does.

4 Relations between random variables

So far we've looked at single random variables. But economics and finance — and lots of other things as well — are concerned not with the properties of single random variables, but with relations among two or more of them. It's not enough to know the distributions of GDP growth and equity returns, we'd also like to know if they're related, and if so, how. We need a language for talking about that.

We start with *independent* random variables, which we define shortly. The setup for "multivariate" (more than one) random variables is similar to what we've seen. The probability density function for a two-dimensional random variable (x_1, x_2) might be expressed $p(x_1, x_2)$. We say x_1 and x_2 are independent if this factors into separate functions: $p(x_1, x_2) = p_1(x_1)p_2(x_2)$. You might think of coin flips. If two flips are independent, then the probability of two heads (say) is just the product of each head separately. If the probability of heads is one-half each time, then the probability of two heads is one-fourth. That's what independence is.

Dependence — the opposite of independence — comes in many forms. The most direct connection is a linear one, which we can document with *covariances* and *correlations*. The covariance between two random variables x_1 and x_2 is

$$\mathrm{Cov}(x_1, x_2) = E[x_1 - E(x_1)][x_2 - E(x_2)].$$

It's an example of a "joint moment," a moment that involves two or more random variables. If the covariance is positive, high values of x_1 are associated, more often than not, with high values of x_2. If negative, the reverse. Their correlation is a scale-free version:

$$\mathrm{Corr}(x_1, x_2) = \frac{\mathrm{Cov}(x_1, x_2)}{\mathrm{Var}(x_1)^{1/2}\mathrm{Var}(x_2)^{1/2}}, \tag{7}$$

a number between minus one and one. You might verify for yourself that if we look instead at the variables $a + bx_1$ and $c + dx_2$, the correlation is the same: a and c drop out when we subtract the means and, except for sign, b and d cancel when we calculate the ratio.

Sample analogs are more or less predictable. The sample means and variances, of course, are

$$\bar{x}_i \;=\; T^{-1}\sum_{t=1}^{T} x_{it}, \quad \mathrm{Var}(x_i) \;=\; T^{-1}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)^2$$

for variables $i = 1, 2$. The sample covariance of x_1 and x_2 is

$$\mathrm{Cov}(x_1, x_2) \;=\; T^{-1}\sum_{t=1}^{T}(x_{1t}-\bar{x}_1)(x_{2t}-\bar{x}_2)$$

and the sample correlation is constructed from the sample variances and covariance by (7). You can generally get the idea from a scatterplot of the two variables. If you're taken by this possibility, search "Anscombe's quartet."

Correlation is a measure of linear association. It's entirely possible for two variables to be related in a nonlinear way but have a zero covariance and correlation. Consider this example: Let x_1 take on the values $\{-1, 0, 1\}$ with probability one-third each. Then let $x_2 = (x_1)^2$. The two variables are clearly related, but their covariance is zero.

5 Sums and mixtures

We'll sometimes create random variables from combinations, either sums or mixtures of independent components. Here's how that works.

Sums. Suppose we start with the independent random variables x_1 and x_2. The distribution of the sum $y = x_1 + x_2$ depends (evidently) on the distributions of x_1 and x_2. The mgf of y is

$$h_y(s) \;=\; E\left(e^{sy}\right) \;=\; E\left(e^{s(x_1+x_2)}\right) \;=\; E\left(e^{sx_1}e^{sx_2}\right) \;=\; E\left(e^{sx_1}\right)E\left(e^{sx_2}\right).$$

The last step follows from independence of x_1 and x_2. It implies $h_y(s) = h_1(s)h_2(s)$: the mgf of the sum is the product of the mgf's of the components.

The cgf's of sums are even simpler. The cgf of the sum is the sum of the cgf's:

$$k_y(s) \;=\; \log h_y(s) \;=\; \log h_1(s) + \log h_2(s) \;=\; k_1(s) + k_2(s). \tag{8}$$

[Think about this a minute. Make sure you follow the notation.] In words: the cgf of a sum of independent random variables is the sum of the cgf's of the components. From this it follows that the cumulants of the sum y are the sums of the cumulants of the components x_1 and x_2. That is: If component x_i has jth cumulant κ_{ij}, then the jth cumulant of $y = x_1 + x_2$ is $\kappa_{1j} + \kappa_{2j}$. [If this isn't clear, stare at (8) and (6).]

We can run through our usual list of examples and see how this works. In each case, let x_1 and x_2 be independent random variables with the same distribution. What is the distribution of $y = x_1 + x_2$?

- *Bernoulli.* The cgf of each x_i is $k_i(s) = \log\left[(1-\omega) + \omega e^s\right]$. The cgf for y is the sum of two of these: $k_y(s) = 2\log\left[(1-\omega) + \omega e^s\right]$. If you look up the cgf for a binomial random variable, you'll see it has this form.

- *Poisson.* The cgf of each x_i is $k_i(s) = \omega(e^s - 1)$, so y has cgf $k_y(s) = 2\omega(e^s - 1)$. Therefore y inherits the Poisson distribution of the components, but the intensity parameter changes to 2ω. This is a useful property; we'll put it to good use when we do option pricing.

- *Normal.* The cgf of each x_i is $k_i(s) = \mu s + \sigma^2 s^2/2$, so $k_y(s) = 2\mu s + 2\sigma^2 s^2/2$. This is normal, too, with double the mean and double the variance (not the standard deviation!).

Thinking ahead, sums are a device for introducing nonnormality: If at least one of the components is nonnormal, then so is the sum. [Can you see why that is?]

Mixtures. A mixture is a more complicated object, but also a more interesting one. A mixture is often defined as a weighted average of probability densities. Suppose we have a bunch of pdf's $p_i(x)$. Then a mixture is a random variable whose pdf is a weighted average of pdf's:

$$p(x) \quad = \quad \sum_i \omega_i p_i(x),$$

with weights $\{\omega_i\}$ that are positive and sum to one. From this we see that p is a legitimate pdf. [Think about it.]

I regard mixing as a device for generate interesting distributions, but we could give it a physical interpretation as a two-stage random variable. First we draw a number i from a distribution with probabilities ω_i. Given a draw for i, we then draw from the distribution of $p_i(x)$. [Draw two-stage event tree.]

Let's see how this works when we mix two distributions: with probability $1 - \omega$ x has pdf $p_1(x)$ and with probability ω it has pdf $p_2(x)$. The overall pdf is therefore

$$p(x) \quad = \quad (1-\omega)p_1(x) + \omega p_2(x).$$

From this it follows that its mgf is also a weighted average:

$$h_y(s) \quad = \quad (1-\omega)E_1\left(e^{sx}\right) + \omega E_2\left(e^{sx}\right) \quad = \quad (1-\omega)h_1(s) + \omega h_2(s),$$

where E_j means the expectation computed from $p_j(x)$. The cgf is the log of this. This gives us the log of a sum, which in some ways is a mess, but we'll see that it produces some interesting properties.

It's less obvious, perhaps, but this is also a device for introducing nonnormality. That's true here even if the components are normal. Some of the most popular models in finance start with mixtures like this. Most applications of (one of several things called) the Merton option pricing model, for example, use Poisson mixtures of normals.

Bottom line

We have seen how to formalize the idea of random variables, aka things that are random. We also have some summary measures that describe it. The mean and variance describe location and dispersion. Skewness and excess kurtosis describe the shape of the distribution and are independent of scale and location. They are useful in identifying departures from normality, the first focusing on asymmetry, the second on the tails. Covariances and correlations describe relations between random variables. We'll use them all — intensively — from now on.

More

This is basic probability theory, which you can find in lots of places. You should think seriously about taking any course with a similar title. One thing to keep in mind: the level of mathematical sophistication ranges from low to extremely high. What we've done is between low and medium, but that's not a bad place to be. The high end typically uses measure theory, an approach to probability that integrates discrete and continuous probability distributions. In some sense it's the right way to do things, but it's not one that shows up much in applied work. As one wag quipped: "A theoretical statistician knows all about measure theory but has never seen a measurement whereas the actual use of measure theory by the applied statistician is a set of measure zero." That said, I've run across practical situations in which measure-theoretic concepts were helpful.

Wikipedia is very good for properties of particular distributions. Search for (say) "Poisson distribution" and you'll find lots of what we've done here, and more. Wikipedia's also ok on the underlying mathematics, but you're never sure what level you'll get: too high, too low, or just right.

Practice problems

1. *Digital option.* A digital option pays some constant amount — say, one hundred — in some situations, nothing in others.
   (a) What set of states do we need to describe this asset's payoffs?
   (b) What random variable connects the payoffs to the states?
   (c) Suppose the probability of a positive payoff is θ. What conditions on θ guarantee that we have a legitimate probability distribution?
   (d) What is the mean payoff?
   (e) What is the variance of the payoff? The standard deviation?
   (f) For what value(s) of θ is the variance largest? Smallest?

   Answer.
   (a) Two states correspond to the two payoffs. We could use, for example, $z = 0$ for no payoff, $z = 1$ for a payoff of one hundred.
   (b) With this choice of states, the payoff is a random variable $x(z) = 100z$.

(c) Probabilities must be positive and sum to one. If they sum to one, we have probabilities $(1-\theta, \theta)$, so θ must be between zero and one.

(d) The mean is $(1-\theta) \cdot 0 + \theta \cdot 100 = \theta \cdot 100$.

(e) The variance is (I'm using the formula for a Bernoulli random variable) $\theta(1-\theta)100^2$. The standard deviation is the square root.

(f) This is largest when $\theta = 1/2$, smallest (zero) when θ is zero or one.

2. *Sample moments.* Consider the following observations of a random variable x: $(2, -1, 4, 3)$. Write a Matlab script to answer the following:

   (a) What are the first two sample moments?

   (b) What is the (sample) mean?

   (c) What is the variance? The standard deviation?

   Answer. Here's a script:

```
x = [2,-1,4,3]'
mu1p = sum(x)/4
mu2p = sum(x.^2)/4
mean = mu1p
variance = mu2p - mu1p^2
var_alt = sum((x-mu1p).^2)/4
stddev = sqrt(variance)
```

   (a) The first sample moment is 2 (the average of the observations), the second is 7.5 (the average of the squared observations).

   (b) The mean is the first one: 2.

   (c) The variance is the second moment minus the square of the first: $7.5 - 2^2 = 3.5$. We could also have subtracted the mean from x and computed the average of these squared deviations, which is again 3.5. [Can you show that both methods give you the same answer?] The standard deviation is the square root: $1.87 = 3.5^{1/2}$.

3. *Normal random variables.* Suppose x is standard normal: that is, normal with $\mu = 0$ and $\sigma = 1$.

   (a) What is its pdf?

   (b) What is its cgf?

   (c) Use the cgf to derive its mean and variance.

   (d) Now consider $y = \mu + \sigma x$. What is its cgf? Use it to derive its mean and variance.

   Answer.

   (a) We've seen this one already: $p(x) = (2\pi)^{-1/2} \exp(-x^2/2)$.

   (b) This one, too: $k(s) = s^2/2$.

   (c) The first derivative gives us the mean. $k^{(1)}(s) = s$. The mean is the value at $s = 0$, which is zero. The second derivative gives us the variance: $k^{(2)}(s) = 1$, so the variance is one.

(d) Apply the definition:

$$k_y(s) \;=\; \log E(e^{sy}) \;=\; \log E(e^{s(\mu+\sigma x)}) \;=\; \log\left[e^{s\mu}E(e^{s\sigma x})\right] \;=\; s\mu + k(s\sigma).$$

That gives us $k_y = s\mu + (s\sigma)^2/2$. Its first and second derivatives give us a mean of μ and a variance of σ^2. This bit of Matlab code does the work:

```
syms s mu sigma                        % defines these as symbols
cgf_x = s^2/2;
cgf_y = s*mu + subs(cgf_x, s, s*sigma);

kappa1 = subs(diff(cgf_y,s,1),s,0)  % mean
kappa2 = subs(diff(cgf_y,s,2),s,0)  % variance
```

4. *Cumulants and moments.* Moments are clearly defined as expectations of powers of random variables, $E(x^j)$. But what are cumulants? That's less clear, but we can connect them to moments using the relation between the mgf $h(s)$ and cgf $k = \log h(s)$.

   (a) What is the jth derivative of $h(s)$ evaluated at $s = 0$?

   (b) How are derivatives of $k(s)$ connected to derivatives of $h(s)$?

   (c) Show that the third cumulant is the same as the third central moment.

   Answer.

   (a) This is the jth (raw) moment μ_j.

   (b) Since $k(s) = \log h(s)$, the derivatives of k are connected to those of h. Eg, $k^{(1)}(s) = h^{(1)}(s)/h(s)$. That gives us a connection between cumulants (derivatives of k) and moments (derivatives of h).

   (c) We need to work through the derivatives one by one:

$$\begin{aligned} k^{(1)}(s) &= h^{(1)}(s)/h(s) \\ k^{(2)}(s) &= [h(s)h^{(2)}(s) - h^{(1)}(s)^2]/h(s)^2 \;=\; h^{(2)}(s)/h(s) - k^{(1)}(s)^2 \\ k^{(3)}(s) &= [h(s)h^{(3)}(s) - h^{(2)}(s)h^{(1)}(s)]/h(s)^2 - 2k^{(1)}(s)k^{(2)}(s) \end{aligned}$$

   Now we evaluate at $s = 0$. Since $h(0) = 1$ (why?), the first three cumulants are

$$\begin{aligned} k^{(1)}(0) &= h^{(1)}(0) \;=\; \mu_1' \\ k^{(2)}(s) &= h^{(2)}(0) - h^{(1)}(0)^2 \;=\; \mu_2' - (\mu_1')^2 \\ k^{(3)}(s) &= h^{(3)}(0) - h^{(2)}(0)h^{(1)}(0)] - 2k^{(1)}(0)k^{(2)}(0) \;=\; \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3. \end{aligned}$$

   It's a little tedious to show, but the final expression is the third central moment: $E(x - \mu_1')^3$. To show this, just multiply out the cube and take expectations. For more along these lines, look up cumulants in Wikipedia.

5. *Three-state "Bernoulli."* Consider a 3-state distribution in which we can explore the impact of high-order cumulants. This is sometimes called a "categorical distribution." Let us say, to be concrete, that the state z takes on the values $\{-1, 0, 1\}$ with probabilities $\{\omega, 1 - 2\omega, \omega\}$. A random variable x is defined by $x(z) = \delta z$.

   (a) Does x have a legitimate probability distribution?

(b) What is the moment generating of x? The cumulant generating function?

(c) What are the mean and variance of x?

(d) What are the measures of skewness and excess kurtosis, γ_1 and γ_2? Under what conditions is γ_2 large?

(e) Is there a value of ω that reproduces the values of γ_1 and γ_2 of a normal random variable?

Answer.

(a) Probabilities are positive (maybe greater than or equal to zero) and sum to one. So we're all set if $0 < \omega < 1/2$ (or weak inequalities if you prefer).

(b) The mgf is

$$h(s) \quad = \quad \omega(e^{-\delta s} + e^{\delta s}) + (1 - 2\omega).$$

The cgf is $k(s) = \log h(s)$.

(c) The mean and variance are

$$\begin{aligned} \kappa_1 &= 0 \\ \kappa_2 &= \delta^2 2\omega. \end{aligned}$$

One way to find them is from the derivatives of the cgf.

(d) Skewness and excess kurtosis are

$$\begin{aligned} \gamma_1 &= 0 \\ \gamma_2 &= 1/(2\omega) - 3. \end{aligned}$$

The first is clear, because the distribution is symmetric. The second is a calculation based on cumulants. Or you could compute the 4th central moment directly,

$$\mu_4 \quad = \quad \omega(-\delta)^4 + (1 - 2\omega) \cdot 0^4 + \omega\delta^4 \quad = \quad 2\omega\delta^4,$$

which implies $\gamma_2 = \mu_4/(\mu_2)^2 - 3 = 1/2\omega - 3$. Evidently γ_2 is large when ω is small.

(e) The normal has $\gamma_1 = \gamma_2 = 0$. We get the first automatically and the second if $p = 1/6$.

Matlab code:

```
syms s omega delta                  % defines these as symbols
mgf = omega*(exp(-s*delta)+exp(s*delta)) + (1-2*omega);
cgf = log(mgf);

kappa1 = subs(diff(cgf,s,1),s,0)    % mean
kappa2 = subs(diff(cgf,s,2),s,0)    % variance
kappa3 = subs(diff(cgf,s,3),s,0)
kappa4 = subs(diff(cgf,s,4),s,0)

gamma1 = kappa3/kappa2^(3/2)
gamma2 = kappa4/kappa2^2
simplify(gamma2)                    % sometimes this helps
```

6. *Exponential random variables.* We say x is exponential if its pdf is $p(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ and $\lambda > 0$. It's a one-sided distribution and (as we shall see) skewed to the right. [Graph it, you'll get the idea.]

   (a) Show that the cgf is

   $$k(s) \quad = \quad -\log(1 - s/\lambda).$$

   (b) Use the cgf to compute the variance, skewness, and excess kurtosis of x. How do they compare to those of normal random variables?

   (c) Suppose we have a random variable $y = a + x$ with a mean of 10 and a standard deviation of 2. What values of a and λ reproduce these values?

   Answer.

   (a) Apply the definition of the mgf:

   $$h(s) \quad = \quad \int_0^\infty \lambda e^{(s-\lambda)x} dx \quad = \quad \lambda(s-\lambda)^{-1} e^{(s-\lambda)x}\Big|_0^\infty \quad = \quad 1/(1 - s/\lambda).$$

   This converges for $s < \lambda$, which includes $s = 0$. The cgf follows from taking the log.

   (b) The first four cumulants are $\kappa_1 = 1/\lambda$, $\kappa_2 = 1/\lambda^2$, $\kappa_3 = 2/\lambda^3$, and $\kappa_4 = 6/\lambda^4$. Skewness and excess kurtosis are therefore $\gamma_1 = 2$ and $\gamma_2 = 6$. Both are zero for normal random variables.

   (c) The mean and standard deviation of y are $a + 1/\lambda$ and $1/\lambda$. We hit the given targets if $\lambda = 1/2$ and $a = 8$.

   Matlab code: adapt from earlier problems.

7. *The sum of normals is normal.* The idea here is to use cumulant generating functions (cgfs) to show that the sum of independent normal random variables is also normal. It's helpful to break the problem into manageable pieces, like this:

   (a) Consider two independent random variables x_1 and x_2, not necessarily normal. Show that the cgf of the sum $y = x_1 + x_2$ is the sum of their cgfs:

   $$k_y(s) \quad = \quad k_1(s) + k_2(s).$$

   Hint: Note the form of the pdf and apply the definition of the cgf.

   (b) Suppose $x_i \sim \mathcal{N}(\kappa_{i1}, \kappa_{i2})$. [This bit of notation means: x_i is normally distributed with mean κ_{i1} and variance κ_{i2}.] What is x_i's cgf?

   (c) Use (a) to find the cgf of $y = x_1 + x_2$, with (x_1, x_2) as described in (b) (namely, normal with given means and variances). How do you know that y is also normal? What are its mean and variance?

   (d) Extend this result to $y = ax_1 + bx_2$ for any real numbers (a, b).

Answer.

(a) Recall that if x_1 and x_2 are independent, their pdf factors: $p_{12}(x_1, x_2) = p_1(x_1)p_2(x_2)$. That means the mgf of $x_1 + x_2$ is the product of their individual mgf's: $h_y(s) = h_1(s)h_2(s)$. We take the log to get the cgf's: $k_y(s) = \log h_y(s) = \log h_1(s) + \log h_2(s) = k_1(s) + k_2(s)$.

(b) $k_i(s) = s\kappa_{1i} + s^2\kappa_{2i}/2$. (If this isn't burned into your memory already, please burn it in now.)

(c) Sum the cgf's:

$$\begin{aligned} k_y(s) &= k_1(s) + k_2(s) \\ &= (s\kappa_{11} + s^2\kappa_{12}/2) + (s\kappa_{21} + s^2\kappa_{22}/2) \\ &= s(\kappa_{11} + \kappa_{21}) + s^2(\kappa_{12} + \kappa_{22})/2. \end{aligned}$$

It's normal because its mgf has the form of a normal random variable: quadratic in s. In fact, we can pick the mean and variance right out of the formula.

(d) Still normal, but with a change in mean and variance:

$$k_y(s) = s(a\kappa_{11} + b\kappa_{12}) + s^2(a^2\kappa_{21} + b^2\kappa_{22})/2.$$

8. *Normal mixtures.* We'll use a Bernoulli mixture of normals to produce random variables that are decidedly not normal. Consider $x \sim \mathcal{N}(0, 1)$ with probability $1 - \omega$ and $x \sim \mathcal{N}(\theta, 1)$ with probability ω. The idea is that x is standard normal most of the time (probability $1-\omega$), but once in a while (probability ω, with ω small) we get a draw from a normal with a different mean (θ). What does this do to the distribution? Let's see.

(a) What is the pdf? What does it look like?

(b) What is x's cgf?

(c) What are the first three cumulants?

(d) How does this differ from a normal random variable?

Suggestion: Let Matlab do most of the work.

Answer.

(a) The pdf is a weighted average of the two normal components:

$$p(x) = (1-\omega)(2\pi)^{-1/2}\exp(-x^2/2) + \omega\,(2\pi)^{-1/2}\exp[-(x-\theta)^2/2].$$

It's a weighted average of two normals.

(b) The cgf is

$$k(s) = \log\left[(1-\omega)e^{s^2/2} + \omega e^{\theta s + s^2/2}\right].$$

No need to derive this, it follows immediately from things we've done already: Bernoulli mixtures and normal mgf's.

(c) Matlab gives us

$$\begin{aligned}\kappa_1 &= \omega\theta \\ \kappa_2 &= \theta^2\omega(1-\omega)+1 \\ \kappa_3 &= \theta^3\omega(1-\omega)(1-2\omega).\end{aligned}$$

Here's the code:

```
syms s theta omega                    % defines these as symbols
mgf = (1-omega)*exp(s^2/2) + omega*exp(theta*s + s^2/2);
cgf = log(mgf);
kappa1 = subs(diff(cgf,s,1),s,0)    % mean
kappa2 = subs(diff(cgf,s,2),s,0)    % variance
kappa3 = subs(diff(cgf,s,3),s,0)
factor(kappa3)                        % sometimes works
```

(d) It's obviously not normal, but the key is the third cumulant κ_3, which is zero for normals but takes the sign of θ here. The rest is similar to the Bernoulli on its own. If you'd like a picture, try this:

```
x = [-4:0.1:4]';

p1 = exp(-x.^2/2)./sqrt(2*pi);
mu = -2; sigma = 1;
p2 = exp(-(x-mu).^2/(2*sigma^2))./sqrt(2*pi*sigma^2);
omega = 0.2;
pmix = (1-omega)*p1 + omega*p2;

plot(x,p1,'b')
hold on
plot(x,pmix,'m')
text(0.4, 0.38, 'blue=std normal, magenta=mixture')
```

9. *Sample covariances and correlations.* We continue an earlier problem, where $x = (2, -1, 4, 3)$, and add $y = (10, -5, 3, 0)$.

(a) What is the (sample) variance of y?

(b) What is the covariance of x and y?

(c) What is the correlation of x and y?

Answer.

(a) The sample variance is 29.5.

(b) The sample covariance is 5.25.

(c) The correlation is

$$\text{Corr}(x,y) = \frac{\text{Cov}(x,x)}{\text{Var}(x)^{1/2}\text{Var}(y)^{1/2}} = \frac{5.25}{3.5^{1/2}29.5^{1/2}} = 0.5167.$$

We noted earlier that correlations are independent of scale. You can verify that by replacing y with $3y + 17$ and redoing the calculations. Sample Matlab code:

```
x = [2,-1,4,3]'
y = [10, -5, 3, 0]'
xbar = mean(x)
ybar = mean(y)
varx = mean((x-xbar).^2)
vary = mean((y-ybar).^2)
covxy = mean((x-xbar).*(y-ybar))
corrxy = covxy/sqrt(varx*vary)
```