Random Variables [1]

Moments: summarizing joint variation for random variables

MOMENTS ARE of a population or probability distribution. *Statistics*, on the other hand, are summaries of a sample. This conceptual distinction is easily forgotten when doing data analysis: for every moment, there is a corresponding statistic with a very similar definition, and this can create confusion. It's therefore worth repeating: moments describe *populations or probability distributions*, while statistics describe *samples*. This set of notes is about moments of random variables, not statistics.

You have already encountered two ways of a summarizing a single random variable: by quoting its expected value and its variance. These quantities describe two important features—the center and dispersion, respectively—of a probability distribution.

These familiar concepts will still be useful, but no longer sufficient, when two or more variables are in play. In this sense, a quantitative relationship is much like a human relationship: you can't describe one by simply listing off facts about the characters involved. You may know that Homer likes donuts, works at the Springfield Nuclear Power Plant, and is fundamentally decent despite being crude, obese, and incompetent. Likewise, you may know that Marge wears her hair in a beehive, despises the *Itchy and Scratchy Show*, and takes an active interest in the local schools. Yet these facts alone tell you little about their marriage. A quantitative relationship is the same way: if you ignore the interactions of the "characters," or individual variables involved, then you will miss the best part of the story.

Moments of one variable

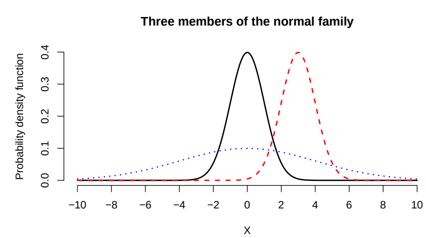
The expected value and variance are just two examples of a general concept called a *moment*. Informally, a moment is a quantitative description of the geometry or shape of a probability distribution. The term is borrowed from physics, where it comes up in many contexts: magnetic moment, electric-dipole moment, moment of inertia, and so forth.

The idea of a moment is best understood with the help of an ice-skating bat spinning in place. By changing the placement of his wings, the bat changes his *moment of inertia*, a one-number summary of how his body mass is distributed in space relative to his axis of rotation. Because angular momentum is conserved, the bat's moment of inertia is inversely related to his rotational velocity. This physical law is exploited to dramatic visual effect in competition, where ice skaters often end a routine by drawing their arms in closer and closer, thereby spinning faster and faster.

Other examples of a moment in physics follow the same conceptual pattern as the moment of inertia: each is a concise geometric description of a body or physical system. The term itself is etymologically related to "momentum;" according to the *Oxford English Dictionary*, it seems to have first been used in this sense in Kater and Lardner's 1830 *Treatise on Mechanics*.[2]

[2] Although Milton came close in 1641: "All the moments and turnings of humane occasions are mov'd to and fro as upon the axle of discipline." (*The reason of church-governement urg'd against prelaty*, X.135).

By metaphorical extension, a moment in probability theory is a geometric description of your uncertainty about a random variable. Recall these three examples of a normal distribution at right, each with a different set of moments. The mean of a random variable X describes where the probability distribution for X is centered; metaphorically, it is like the skater's center of gravity. The variance describes how dispersed the distribution is around its center; metaphorically, it is like the spread of the skater's arms.



Moments in physics have precise mathematical definitions. For example, if a system comprises n different masses $m_1, \ldots, m_n$ placed at radiuses $r_1, \ldots, r_n$ around a common rotational axis, then the system's moment of inertia is

$$I = \sum_{i=1}^{n} m_i r_i^2 \, .$$

In staring at any formula like this, the important question to ask yourself is: *how does the math formalize the intuition?* Here, it does so straightforwardly. The further the objects are from the center, the larger the values of r_i, and therefore the larger the moment of inertia. The quantity I is a summary of the geometric distribution of mass in the system. It doesn't tell you everything about the system, but it does tell you something useful: does the mass tend to concentrate near the axis of rotation, or far away from it?

Moments in probability theory also have precise mathematical definitions. Suppose that X is a discrete random variable that takes on values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$, respectively. The kth moment of X is defined as the expected value of the kth power of X, or

$$E(X^k) = \sum_{i=1}^{n} p_i x_i^k \, .$$

For continuous random variables, there is calculus-based version of the formula:

$$E(X^k) = \int_\Omega x^k p(x) \, dx \, ,$$

where $p(x)$ is the *probability density function* (or p.d.f.) of the random variable X, and Ω is the space of all possible values that X might take on. If you compute the area between two points under the curve of the probability density function, you will get the probability that the random variable will take on a value between those two points.

There is an striking correspondence between this and the formula for the moment of inertia: the probabilities p_i are like the masses m_i, while the values x_i are like the radiuses r_i. In fact, the analogy with physical mass is so instructive that, when we describe a probability distribution by listing the possible values x_i together with their probabilities p_i, we are said to be specifying the *probability mass function* of the distribution.

Likewise, the kth *central moment* of X is

$$E\left[\{X - E(X)\}^k\right] = \sum_{i=1}^{n} p_i \{x_i - E(X)\}^k \, .$$

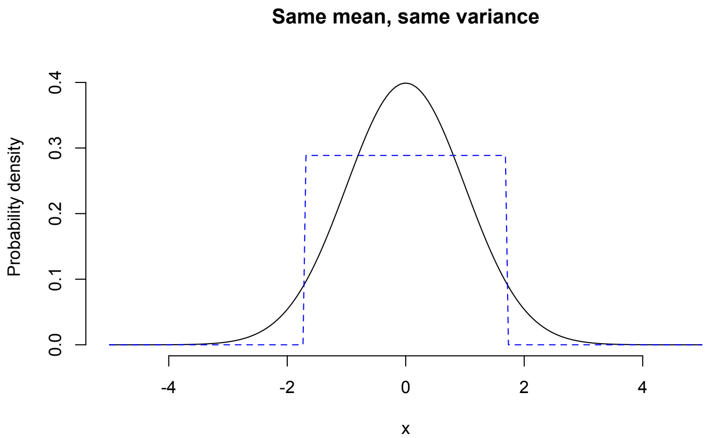
The mean of a probability distribution is its first moment; the variance is its second central moment. Higher-order moments also have geometric interpretations. For example, a probability distribution's skewness (or lopsidedness) is measured by the third moment, while its tail weight (or propensity to produce extreme events) is measured by the fourth moment.

The following two points about moments are worth remembering.

1. Moments are merely summaries. Two probability distributions can have the same mean and the same variance, and yet be very different:

   With few exceptions, the only way to perfectly characterize an entire distribution is to quote the probability mass function—or, for a continuous random variable, the probability density function.

2. Like everything in mathematics, the definition of a moment is just a human convention, agreed upon by a body of working scientists and statisticians. There is nothing holy about this



definition; it just happens to be one that conveys information which people find useful.

Joint distributions and the covariance of two random variables

A moment summarizes the probability distribution for one variable. To summarize relationships between more than one variable, we will appeal to the concept of a *mixed moment*, which summarizes a joint distribution.

A *joint distribution* is an exhaustive list of joint outcomes for two or more variables at once, together with the probabilities for each of these outcomes. For example, the table below depicts a simple, stylized joint distribution for the rain and average wind speed on a random day in February.

Outcome	Wind (mph)	Rain (inches)	Probability
1	5	1	0.4
2	5	3	0.1
3	15	1	0.1
4	15	3	0.4

In this simple case, each variable can take one of only two values, and so there are only four possible joint outcomes, whose probabilities must sum to 1. Notice, too, that the joint distribution depicts a positive relationship between wind and rain: when one is high, the other tends to be high as well.

To quantify this relationship, define the *covariance* of two random variables X and Y as

$$\text{cov}(X,Y) = E\Big\{[X - E(X)][Y - E(Y)]\Big\} = \sum_{i=1}^{n} p_i \left[x_i - E(X)\right]\left[y_i - E(Y)\right].$$

This sum is over all possible joint outcomes for X and Y. In the wind/rain example, the expected values for wind speed (X) and rainfall (Y) are

$$\begin{aligned} E(X) &= \sum_{i=1}^{n} p_i x_i = 0.4 \cdot 5 + 0.1 \cdot 5 + 0.1 \cdot 15 + 0.4 \cdot 15 = 10 \\ E(Y) &= \sum_{i=1}^{n} p_i y_i = 0.4 \cdot 1 + 0.1 \cdot 1 + 0.1 \cdot 3 + 0.4 \cdot 3 = 2 \end{aligned}$$

Plugging these numbers into the formula for covariance, we get

$$\begin{aligned} \text{cov}(X,Y) &= E\Big\{[X - E(X)][Y - E(Y)]\Big\} \\ &= 0.4 \cdot (5-10)(1-2) + 0.1 \cdot (5-10)(3-2) + 0.1 \cdot (15-10)(1-2) + 0.4 \cdot (15-10)(3-2) \\ &= 0.4 \cdot (5) + 0.1 \cdot (-5) + 0.1 \cdot (-5) + 0.4 \cdot (5) \\ &= 3\,. \end{aligned}$$

Again, ask yourself: *how does the mathematical definition of covariance formalize the intuition behind the concept of dependence?* Try reasoning through the formula, and its application to this example, on your own.

You may notice the following: in the third line of the above computation, the positive terms correspond to joint outcomes when wind speed and rainfall are on the *same side* of their respective means—that is, both above the mean, or both below it. The negative terms, on the other hand, correspond to outcomes where the two quantities are on *opposite sides* of their respective means. In this case, the "same side" outcomes are more likely than the "opposite side" outcomes, and therefore the covariance is positive.

Correlation as standardized covariance

The covariance is our first example of a mixed moment. It provides one way of quantifying the direction and magnitude of association between two random variables X and Y.

One difficulty that arises in interpreting covariance, however, is that it depends upon the scale of measurement for the two sets of observations. For example, suppose we measured rain in millimeters, rather than inches, as in the following table.

Outcome	Wind (mph)	Rain (mm)	Probability
1	5	25.4	0.4
2	5	76.2	0.1
3	15	25.4	0.1
4	15	76.2	0.4

Now $E(Y) = 50.8$, and the wind and rain variables have covariance

$$\begin{aligned}\mathrm{cov}(X,Y) &= 0.4 \cdot (5-10)(-25.4) + 0.1 \cdot (5-10)(25.4) + 0.1 \cdot (15-10)(-25.4) + 0.4 \cdot (15-10)(25.4) \\ &= 76.2\,.\end{aligned}$$

This is 25.4 times as big as 3, the answer from before. And yet we wouldn't say that wind and rain are 25.4 times as "dependent" as they were before; the new numbers describe exactly the same probability distribution, just in different units. Clearly we need a measure of dependence that is invariant to changes in scale. (Interestingly, 25.4 is precisely the number of millimeters in a single inch, a fact which might suggest to you how covariances behave when you multiply one of the variables by a constant. More on that later.)

One such scale-invariant measure is Pearson's *product-moment correlation coefficient*, often called simply the correlation coefficient. (There are other kinds of correlation coefficients as well, and so sometimes we must distinguish them from one another.) The Pearson coefficient, named after English statistician Karl Pearson, is on a standardized scale running from -1 (perfect negative correlation) to $+1$ (perfect positive correlation).

The Pearson correlation coefficient for two random variables X and Y is just their covariance, rescaled by their respective variances:

$$\mathrm{cor}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)} \cdot \sqrt{\mathrm{var}(Y)}}\,.$$

Let's apply this definition to joint distribution for wind speed and rain fall measured in inches:

$$\mathrm{cor}(X,Y) = \frac{3}{\sqrt{25} \cdot \sqrt{1}} = 0.6\,.$$

And with rain measured in millimeters,

$$\mathrm{cor}(X,Y) = \frac{76.2}{\sqrt{25} \cdot \sqrt{645.16}} = 0.6\,.$$

There is a common factor of 25.4 that appears in both the numerator and denominator. It cancels, leaving us with a scale-invariant quantity. If the correlation between two variables is 0, then they are said to be uncorrelated.[3]

[3] Although not necessarily independent!

Functions of random variables

A VERY IMPORTANT set of equations in probability theory describes what happens when you construct a new random variable as a linear combination of other random variables—that is, when

$$W = aX + bY + c$$

for some random variables X and Y and some constants a, b, and c.

The fundamental question here is: how does *joint* variation in X and Y (that is, correlation) influence the behavior a random variable formed by adding X and Y together? To jump straight to the point, it turns out that

$$\begin{aligned} E(W) &= aE(X) + bE(Y) + c && (1) \\ \text{var}(W) &= a^2\,\text{var}(X) + b^2\,\text{var}(Y) + 2ab\,\text{cov}(X,Y)\,. && (2) \end{aligned}$$

Why would you care about a linear combination of random variables? Consider a few examples:

- You know the distribution for X, the number of points a basketball team will score in one quarter of play. Then the random variable describing the points the team will score in four quarters of play is $W = 4x$.
- A weather forecaster specifies a probability distribution for tomorrow's temperature in Celsius (a random variable, C). You can compute the moments of C, but you want to convert to Fahrenheit (another random variable, F). Then F is also a random variable, and is a linear combination of the one you already know: $F = (9/5)C + 32$.
- You know the joint distribution describing your uncertainty as to the future prices of two stocks X and Y. A portfolio of stocks is a linear combination of the two; if you buy 100 shares of the first and 200 of the second, then

  $$W = 100X + 200Y$$

  is a random variable describing the value of your portfolio.

- Your future grade on the statistics midterm is X_1, and your future grade on the final is X_2. You describe your uncertainty for these two random variables with some joint distribution. If the midterm counts 40% and the final 60%, then your final course grade is the random variable

$$C = 0.4X_1 + 0.6X_2\,,$$

  a linear combination of your midterm and final grades.

- The speed of Rafael Nadal's slice serve is a random variable S_1. The speed on his flat serve is S_2. If Rafa hits 70% slice serves, his opponent should anticipate a random service speed equal to $0.7S_1 + 0.3S_2$.

In all five cases, it is useful to express the moments of the new random variable in terms of the moments of the original ones. This saves you a lot of calculational headaches! We'll now go through the mathematics of deriving Equations (1) and (2).

Multiplying a random variable by a constant

Let's first examine what happens when you make a new random variable W by multiplying some other random variable X by a constant:

$$W = aX\,.$$

This expression means that, whenever $X = x$, we have $W = ax$. Therefore, if X takes on values $x_1, \ldots, x_n$ with probability $p_1, \ldots, p_n$, then we know that

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

and so

$$E(W) = \sum_{i=1}^{n} a x_i p_i = a \sum_{i=1}^{n} x_i p_i = aE(X)\,.$$

The constant a simply comes out in front of the original expected value. Mathematically speaking, this means that the expectation is a linear function of a random variable.

The variance of W can be calculated in the same way. By definition,

$$\mathrm{var}(X) = \sum_{i=1}^{n} p_i \{x_i - E(X)\}^2\,.$$

Therefore,

$$\begin{aligned}
\text{var}(W) &= \sum_{i=1}^{n} p_i \{ax_i - E(W)\}^2 \\
&= \sum_{i=1}^{n} p_i \{ax_i - aE(X)\}^2 \\
&= \sum_{i=1}^{n} p_i a^2 \{x_i - E(X)\}^2 \\
&= a^2 \sum_{i=1}^{n} p_i \{x_i - E(X)\}^2 \\
&= a^2 \text{ var}(X)
\end{aligned}$$

Now we have a factor of a^2 out front.

What if, in addition to multiplying X by a constant a, we also add another constant c to the result? This would give us

$$W = aX + c\,.$$

To calculate the moments of this random variable, revisit the above derivations on your own, adding in a constant term of c where appropriate. You'll soon convince yourself that

$$\begin{aligned}
E(W) &= aE(X) + c \\
\text{var}(W) &= a^2 \text{var}(X)\,.
\end{aligned}$$

The constant simply gets added to the expected value, but doesn't change the variance at all.

A linear combination of two random variables

Suppose X and Y are two random variables, and we define a new random variable as $W = aX + bY$ for real numbers a and b. Then

$$\begin{aligned}
E(W) &= \sum_{i=1}^{n} p_i \{ax_i + by_i\} \\
&= \sum_{i=1}^{n} p_i a x_i + \sum_{i=1}^{n} p_i b y_i \\
&= a \sum_{i=1}^{n} p_i x_i + b \sum_{i=1}^{n} p_i y_i \\
&= aE(X) + b(E(Y)\,.
\end{aligned}$$

Again, the expectation operator is linear.

The variance of W, however, takes a bit more algebra:

$$\begin{aligned}
\mathrm{var}(W) &= \sum_{i=1}^{n} p_i \Big\{ [ax_i + by_i] - [aE(X) + bE(Y)] \Big\}^2 \\
&= \sum_{i=1}^{n} p_i \Big\{ [ax_i - aE(X)] + [by_i - bE(Y)] \Big\}^2 \\
&= \sum_{i=1}^{n} p_i \Big\{ [ax_i - aE(X)]^2 + [by_i - bE(Y)]^2 + 2ab[x_i - E(X)][y_i - E(Y)] \Big\} \\
&= \sum_{i=1}^{n} p_i [ax_i - aE(X)]^2 + \sum_{i=1}^{n} p_i [by_i - bE(Y)]^2 + \sum_{i=1}^{n} p_i 2ab[x_i - E(X)][y_i - E(Y)] \\
&= \mathrm{var}(aX) + \mathrm{var}(bY) + 2ab\mathrm{cov}(X,Y) \\
&= a^2\mathrm{var}(X) + b^2\mathrm{var}(Y) + 2ab\mathrm{cov}(X,Y)
\end{aligned}$$

The covariance of X and Y strongly influences the variance of their linear combination. If the covariance is positive, then the variance of the linear combination is *more than* the sum of the two individual variances. If the covariance is negative, then the variance of the linear combination is *less than* the sum of the two individual variances.

An example: portfolio choice under risk aversion

Let's revisit the portfolio-choice problem posed above. Say you plan to allocate half your money to one asset X, and the other half to some different asset Y. Look at Equations (1) and (2), which specify the expected value and variance of your portfolio in terms of the moments of the joint distribution for X and Y. If you are a risk-averse investor, would you prefer to hold two assets with a positive covariance or a negative covariance?

x	y	$P(x,y)$
1.0	1.0	0.15
1.0	1.1	0.10
1.0	1.2	0.05
1.1	1.0	0.10
1.1	1.1	0.20
1.1	1.2	0.10
1.2	1.0	0.05
1.2	1.1	0.10
1.2	1.2	0.15

Table 1: Positive covariance.

To make things concrete, let's imagine that the joint distribution for X and Y is given in the table at right. Each row is a possible joint outcome for X and Y: the first column lists the possible values of X; the second, the possible values of Y; and the third, the probabilities for each joint outcome. You should interpret the numbers in the X and Y columns as the value of \$1 at the end of the investment period—for example, after one year. If $X = 1.1$ after a year, then your holdings of that stock gained 10% in value.

Under this joint distribution, a single dollar invested in a portfolio with a 50/50 allocation between X and Y is a random variable

W. This random variable has an expected value of 1.1 and variance

$$\begin{aligned} \mathrm{var}(W) &= 0.5^2 \mathrm{var}(X) + 0.5^2 \mathrm{var}(Y) + 2 \cdot 0.5^2 \cdot \mathrm{cov}(X, Y) \\ &= 0.5^2 \cdot 0.006 + 0.5^2 \cdot 0.006 + 2 \cdot 0.5^2 \cdot (0.002) \\ &= 0.004 \,, \end{aligned}$$

for a standard deviation of $\sqrt{0.004}$, or about 6.3%.

What if, on the other hand, the asset returns were negatively correlated, as they are in the table at right? (Notice which entries have been switched around, compared to the previous joint distribution.)

x	y	$P(x,y)$
1.0	1.0	0.05
1.0	1.1	0.10
1.0	1.2	0.15
1.1	1.0	0.10
1.1	1.1	0.20
1.1	1.2	0.10
1.2	1.0	0.15
1.2	1.1	0.10
1.2	1.2	0.05

Table 2: Negative covariance.

Under this new joint distribution, the expected value of $1 invested in a 50/50 portfolio is still 1.1. But since the covariance between X and Y is now negative, the variance of the portfolio changes:

$$\begin{aligned} \mathrm{var}(W) &= 0.5^2 \mathrm{var}(X) + 0.5^2 \mathrm{var}(Y) + 2 \cdot 0.5^2 \cdot \mathrm{cov}(X, Y) \\ &= 0.5^2 \cdot 0.006 + 0.5^2 \cdot 0.006 + 2 \cdot 0.5^2 \cdot (-0.002) \\ &= 0.002 \,, \end{aligned}$$

for a standard deviation of $\sqrt{0.002}$, or about 4.5%. Same expected return, but lower variance, and therefore more attractive to a risk-averse investor!

What's going on here? Intuitively, under the first portfolio, where X and Y are positively correlated, the bad days for X and Y tend to occur together. So do the good days. (When it rains, it pours; when it's sunny, it's 100 degrees.) But under the second portfolio, where X and Y are negatively correlated, the bad days and good days tend to cancel each other out. This results in a lower overall level of risk.

The morals of the story are:

1. Correlation creates extra variance.
2. Diversify! (Extra variance hurts your compounded rate of return.)