Combinations of random variables

Linear combinations of random variables

We just reviewed what happens when we shift or scale a data set by a constant value. But now we want to look at what happens when we combine two data sets, either by adding them or subtracting them.

For example, let's say I have two variables: how much time I spend each day walking and biking, W and B respectively. And let's say that I have data on my walking and biking habits for a full year, and I've already found the mean and standard deviation for both variables.

$$\mu_W = 1.1 \qquad\qquad \sigma_W = 0.2$$

$$\mu_B = 0.6 \qquad\qquad \sigma_B = 0.1$$

But now I want to know the mean for the sum of my walking and biking time together. In other words, I spend some time walking and biking each day, so I'd like to get an average for my total daily activity time.

We'll call the total activity A, which means we're looking for μ_A (or the expected value $E(A)$). We know that $A = W + B$. Here's the rule we need to remember: when we want to find the mean of the sum, we just find the sum of the means. So because $A = W + B$,

$$\mu_A = \mu_W + \mu_B$$

$$\mu_A = 1.1 + 0.6$$

$$\mu_A = 1.7$$

But if we want to find the standard deviation of these two variables, we can't simply add the standard deviations together. In other words, $\sigma_A = \sigma_W + \sigma_B$ is not a valid equation.

Instead, to find the standard deviation of the total activity, we need to square the two standard deviations. Remember that this is really giving us the variation for both walking and biking.

$$\sigma_W^2 = 0.2^2 = 0.04$$

$$\sigma_B^2 = 0.1^2 = 0.01$$

Then we add these together to get the sum of the variances, which gives us the variance for total activity.

$$\sigma_A^2 = \sigma_W^2 + \sigma_B^2$$

$$\sigma_A^2 = 0.04 + 0.01$$

$$\sigma_A^2 = 0.05$$

Then to find standard deviation for total activity, we take the square root of both sides.

$$\sqrt{\sigma_A^2} = \sqrt{0.05}$$

$$\sigma_A \approx 0.22$$

Instead of the sum, we could also find the difference in my walking and biking times. We could define a new variable for the difference and call it

D. Then the difference is $D = W - B$, and the expected value of the difference would be

$$E(D) = \mu_D = \mu_W - \mu_B$$

$$E(D) = \mu_D = 1.1 - 0.6$$

$$E(D) = \mu_D = 0.5$$

And the standard deviation of the difference would be

$$\sqrt{\sigma_D^2} = \sqrt{\sigma_W^2 + \sigma_B^2}$$

$$\sqrt{\sigma_D^2} = \sqrt{0.04 + 0.01}$$

$$\sqrt{\sigma_D^2} = \sqrt{0.05}$$

$$\sigma_D \approx 0.22$$

One important thing to note is that, regardless of whether we're finding the sum of the variables, or the difference of the variables, in both cases we take the sum of the variances $\sigma_W^2 + \sigma_B^2$. We don't use the sum of the variances for the sum, and the difference of the variances for the difference; we always use the sum for both.

When we find the mean of the sum or difference of variables, it doesn't matter whether or not the variables are dependent or independent. In other words, if the variables are dependent, we can find a valid mean of their sum or difference. And if the variables are independent, we can find a valid mean of their sum or difference.

But in order to find the standard deviation of the sum or difference of two variables, the variables must be independent. So we can summarize what we know about the formulas this way:

	Combination	Mean	Variance
Sum	$S = X + Y$	$\mu_S = \mu_X + \mu_Y$	$\sigma_S^2 = \sigma_X^2 + \sigma_Y^2$
Difference	$D = X - Y$	$\mu_D = \mu_X - \mu_Y$	$\sigma_D^2 = \sigma_X^2 + \sigma_Y^2$

Combinations of normally distributed variables

When we combine variables that are both normally distributed, the combination will be normally distributed as well.

So if we're given the mean and standard deviation of two normally distributed variables, we can calculate the mean and standard deviation of the new combination.

But then, since the combination is normally distributed, we can use what we know about the probability under normal distributions to answer probability questions about the combination.

Example

A popcorn company fills each of its variety popcorn tins with three flavors of popcorn: white cheddar, caramel, and chocolate covered. The amount of each flavor of popcorn that gets packed in the tin is normally distributed with a mean of 1 pound and a standard deviation of 0.1

pounds. The amount of each popcorn flavor is independent from the other flavors.

If W is the total weight of popcorn in a randomly selected tin, find the probability that the tin contains less than 3.25 pounds.

We have three normally distributed variables, one for each flavor. Their means are

White cheddar $\mu_D = 1$

Caramel $\mu_M = 1$

Chocolate covered $\mu_C = 1$

Therefore, the mean of the combination (the mean weight of a full tin) is

$\mu_W = \mu_D + \mu_M + \mu_C$

$\mu_W = 1 + 1 + 1$

$\mu_W = 3$

The standard deviations of the three normally distributed variables are

White cheddar $\sigma_D = 0.1$

Caramel $\sigma_M = 0.1$

Chocolate covered $\sigma_C = 0.1$

To find the standard deviation of the combination (the standard deviation of the weight of a full tin), we'll find the variance of the combination.

$$\sigma_W^2 = \sigma_D^2 + \sigma_M^2 + \sigma_C^2$$

$$\sigma_W^2 = 0.1^2 + 0.1^2 + 0.1^2$$

$$\sigma_W^2 = 0.01 + 0.01 + 0.01$$

$$\sigma_W^2 = 0.03$$

So the standard deviation is

$$\sigma_W = \sqrt{0.03}$$

$$\sigma_W \approx 0.1732$$

Now that we have the mean $\mu_W = 3$ and standard deviation $\sigma_W \approx 0.1732$ of the normally distributed weight of the full tin, we can answer probability questions about the combined normal distribution. We want to find the probability that the tin contains less than 3.25 pounds, so we'll calculate the z-score for 3.25.

$$z = \frac{x - \mu_W}{\sigma_W}$$

$$z = \frac{3.25 - 3}{0.1732}$$

$$z \approx 1.44$$

If we look up $z = 1.44$ in a z-table, we find the value 0.9251, which means there's an approximately 93% chance that the weight of the full tin is less than 3.25 pounds.