Calculating the Change in Talent Variance over Time

-Adam Dorhauer

In my article "Regression with Changing Talent Levels: the Effects of Variance" on the Hardball Times, I talk about how changes in players' true talent levels from day to day reduce the variance of talent in the population overall over time. In other words, the spread in talent over a 100-game sample will be smaller than the spread in talent over a one-game sample. In the article, I gave the following formula to calculate how much the spread in talent is reduced, which I will further explain here:

n ; number of days in the sample
r ; day-to-day correlation for true talent

$var_{t_{0:n-1}}$; variance in talent over n days
var_{t_0} ; variance in talent over one day

$$\frac{var_{t_{0:n-1}}}{var_{t_0}} = \frac{n(1-r^2) - 2r(1-r^n)}{n^2(1-r)^2}$$

**Note: in the THT article, I used d for the number of days instead of n to avoid confusion with another formula that was referenced from a previous article, which used n for something else. For this article, I'm just going to use n for the number of days.*

The value given by the formula is the ratio of talent variance over n days to the talent variance for a single day. In other words, the variance in talent drops by a multiplicative factor that is dependent on the length of the sample and the correlation of talent from day to day.

Now, how do we get that formula?

If we only have two days in our sample, it is not too difficult to calculate the drop in talent variance. Let t_0 be a variable representing player talent levels on Day 1, and t_1 be a variable representing player talent levels on Day 2. We want to find the variance of the average talent levels over both days, or $(t_0+t_1)/2$.

The following formula gives us the variance of the sum of two variables:

$$var_{t_0+t_1} = var_{t_0} + var_{t_1} + 2\, cov_{t_0,t_1}$$

The covariance is directly proportional to the correlation between the two variables and is defined as follows:

$$cov_{t_0,t_1} = sd_{t_0} sd_{t_1} r = var_{t_0} r$$
$$var_{t_0+t_1} = 2\, var_{t_0} + 2\, var_{t_0}\, r = 2\, var_{t_0}(1+r)$$

(Note that $sd_{t_0} sd_{t_1} = var_{t_0} = var_{t_1}$ because the standard deviation and variance for both variables are the same.)

Before we continue, there is an important thing to note. Because we are trying to derive a formula for a ratio (variance in talent over n days divided by variance in talent over one day), we don't necessarily need to calculate the numerator and denominator of that ratio exactly. As long as we can calculate values that are proportional to those values by the same factor, the ratio will be preserved.

Technically, we want the variance of the value $(t_0+t_1)/2$ and not just t_0+t_1, which would be $var_t(1+r)/2$ instead of $2var_t(1+r)$. However, those two values are proportional, so it doesn't really matter for now which we calculate as long as we can also calculate a value for the denominator that is proportional by the same factor.

For two days, the above calculations are simple enough. Once you start adding more days, however, it starts to get more complicated. Fortunately, the above math can also be expressed with a covariance matrix:

	t_0	t_1
t_0	var_0	$cov_{0,1}$
t_1	$cov_{0,1}$	var_1

The variance of the sum t_0+t_1 is equal to the sum of the terms in the covariance matrix, which you can see just gives us the formula: $var_{t_0+t_1} = var_{t_0} + var_{t_1} +$

2 cov_{t_0,t_1}. The covariance matrix is convenient because it can be expanded for any number of days:

Covariance matrix between talent n days apart

	t_0	t_1	t_2	t_3	...	t_{n-1}
t_0	var_0	$cov_{0,1}$	$cov_{0,2}$	$cov_{0,3}$	...	$cov_{0,n-1}$
t_1	$cov_{0,1}$	var_1	$cov_{1,2}$	$cov_{1,3}$	...	$cov_{1,n-1}$
t_2	$cov_{0,2}$	$cov_{1,2}$	var_2	$cov_{2,3}$	...	$cov_{2,n-1}$
t_3	$cov_{0,3}$	$cov_{1,3}$	$cov_{2,3}$	var_3	...	$cov_{3,n-1}$
⋮	⋮	⋮	⋮	⋮	⋱	⋮
t_{n-1}	$cov_{0,n-1}$	$cov_{1,n-1}$	$cov_{2,n-1}$	$cov_{3,n-1}$	...	var_{n-1}

We can also construct a correlation matrix. Given that we know the correlation of talent from one day to the next, this isn't that difficult. If the correlation between talent levels on Day 1 and Day 2 is r, and the correlation between talent levels on Day 2 and Day 3 is also r, we can chain those two facts together to find that the correlation between talent levels on Day 1 and Day 3 is r^2.

The same logic can be extended for any number of days, so that the correlation between talent levels n days apart is r^n:

Correlation matrix between talent n days apart

	t_0	t_1	t_2	t_3	...	t_{n-1}
t_0	r^0	r^1	r^2	r^3	...	r^{n-1}
t_1	r^1	r^0	r^1	r^2	...	r^{n-2}
t_2	r^2	r^1	r^0	r^1	...	r^{n-3}
t_3	r^3	r^2	r^1	r^0	...	r^{n-4}
⋮	⋮	⋮	⋮	⋮	⋱	⋮
t_{n-1}	r^{n-1}	r^{n-2}	r^{n-3}	r^{n-4}	...	r^0

This matrix is more useful than the covariance matrix, because all we need to know to fill in the entire correlation matrix is the value of r. And because correlation is proportional to covariance ($cov_{t_0,t_1} = var_{t_0} r$), the sum of the correlation matrix is proportional to the sum of the covariance matrix.

Our next step, then, is to calculate the sum of the correlation matrix. Notice that the terms on each diagonal going from the top left to bottom right are identical:

	t_0	t_1	t_2	t_3	...
t_0	r^0	r^1	r^2	r^3	...
t_1	r^1	r^0	r^1	r^2	...
t_2	r^2	r^1	r^0	r^1	...
t_3	r^3	r^2	r^1	r^0	...
⋮	⋮	⋮	⋮	⋮	⋱

We can use this pattern to simplify the sum. Since the matrix is symmetrical, we can ignore the terms below the long diagonal and calculate the sum for just the top half of the matrix, and then double it later:

$$
\begin{array}{ccccccccc}
r^0 & r^1 & r^2 & r^3 & \cdots & r^{n-1} & \rightarrow & r^{n-1} \\
 & r^0 & r^1 & r^2 & \ddots & \vdots & & \vdots \\
 & & r^0 & r^1 & \ddots & r^3 & \rightarrow & (n-3)r^3 \\
 & & & r^0 & \ddots & r^2 & \rightarrow & (n-2)r^2 \\
 & & & & \ddots & r^1 & \rightarrow & (n-1)r^1 \\
 & & & & & r^0 & \rightarrow & nr^0
\end{array}
$$

There is one r^0 term in each column of the matrix, so there are n r^0 terms in the sum. Likewise, there are (n-1) r^1 terms, (n-2) r^2 terms, etc. If we group each diagonal into its own distinct term, we get a sum whose terms follow the pattern (n-1)*r^i:

$$\sum_{i=0}^{n-1}(n-i)r^i$$

Applying the distributive property and separating the terms of the sum, we get the following:

$$\sum_{i=0}^{n-1}(nr^i - ir^i)$$

$$n\sum_{i=0}^{n-1} r^i \; - \sum_{i=0}^{n-1} ir^i$$

The first sum is a simple geometric series, which we can calculate using the formula for geometric series:

$$n\sum_{i=0}^{n-1} r^i \; = \; n\frac{(1-r^n)}{(1-r)}$$

The second sum is similar, but the additional i factor makes it a bit trickier since it is no longer a geometric series. We can, however, transform it into a geometric series using a trick where we convert this from a single sum to a double sum, where we replace the expression inside the sum with another sum.

The idea is that each term of the series is itself a separate sum which has i terms of r^i. This sum can be written as follows:

$$ir^i \; = \sum_{h=0}^{i-1} r^i$$

Notice that we switched to using the index h rather than i. This means there is nothing inside the sum that increments on each successive term, and the i acts as a

static value. In other words, this is just adding up the value r^i i times, which is of course equal to ir^i.

$$\sum_{i=0}^{n-1} ir^i = \sum_{i=0}^{(n-1)} \sum_{h=0}^{(i-1)} r^i$$

In order to visualize how this double sum works, we can write down the terms of the sum in an array with i rows and h columns, where the value corresponding to each pair of (i,h) values is r^i. For example, here is what the array would look like with n=4:

	h=0	*h=1*	*h=2*	*h=3*
i=0	r^0	r^0	r^0	r^0
i=1	r^1	r^1	r^1	r^1
i=2	r^2	r^2	r^2	r^2
i=3	r^3	r^3	r^3	r^3

The greyed-out values are included to complete the array, but are not actually part of the sum. If we go through the sum iteratively, we start at i=0, and take the sum of r^i from h=0 to h=-1. Since you can't count up from 0 to -1, there are no values to count in this row, which represents the fact that $ir^i = 0$ when i=0.

Next, we go to i=1, and fill in the values r^1 for k=0 to k=0. The next row, when i=2, we go from h=0 to h=1. And so on.

We are currently taking the sum of each row and then adding those individual sums together. However, we could also start by taking the sum of each column, which would be equivalent to reversing the order of the two sums in our double series:

$$\sum_{i=0}^{(n-1)} \sum_{h=0}^{(i-1)} r^i = \sum_{h=0}^{(n-1)} \sum_{i=h+1}^{(n-1)} r^i$$

Note that the inner sum now goes from i=h+1 to i=n-1, which you can see in the columns of the array of terms above.

This is useful because each column of the array is a geometric series, meaning it will be easy to compute. The sum of each column is just the geometric series from i=0 to i=n-1. Then, to eliminate the greyed-out values from the sum, we subtract the geometric series from i=0 to i=h.

$$\sum_{i=h+1}^{(n-1)} r^i = \frac{1-r^n}{1-r} - \frac{1-r^{h+1}}{1-r} = \frac{r^{h+1} - r^n}{1-r}$$

This is the value for our inner sum, so we plug that back into the outer sum:

$$\sum_{h=0}^{(n-1)} \frac{r^{h+1} - r^n}{1-r}$$

$$\sum_{h=0}^{(n-1)} \frac{r^{h+1}}{1-r} - \sum_{h=0}^{(n-1)} \frac{r^n}{1-r}$$

$$\frac{r}{1-r} \sum_{h=0}^{(n-1)} r^h - \frac{nr^n}{1-r}$$

$$\frac{r(1-r^n)}{(1-r)^2} - \frac{nr^n}{1-r}$$

We now have values for both halves of our original sum, so next we combine them to get the full value:

$$n\sum_{i=0}^{n-1} r^i = \frac{n(1-r^n)}{1-r}$$

$$\sum_{i=0}^{n-1} ir^i = \frac{r(1-r^n)}{(1-r)^2} - \frac{nr^n}{1-r}$$

$$n\sum_{i=0}^{n-1} r^i - \sum_{i=0}^{n-1} ir^i$$

$$= \frac{n(1-r^n)}{1-r} - \left(\frac{r(1-r^n)}{(1-r)^2} - \frac{nr^n}{1-r}\right)$$

$$\frac{n-nr^n}{1-r} + \frac{nr^n}{1-r} - \frac{r(1-r^n)}{(1-r)^2}$$

$$\frac{n}{1-r} - \frac{r(1-r^n)}{(1-r)^2}$$

$$\frac{n(1-r) - r(1-r^n)}{(1-r)^2}$$

We still have one more step to go to calculate the full sum of the correlation matrix. Recall that when we started, we were working with a symmetrical correlation matrix, and because the matrix was symmetrical along the diameter, we set out to find the sum for only the upper half of the matrix. In order to get the sum of the full matrix, we have to double this value:

$$2\left(\frac{n(1-r)\ -\ r(1-r^n)}{(1-r)^2}\right)$$

Finally, note that the long diagonal of the correlation matrix only occurs once in the matrix, so by doubling our initial sum, we are double-counting that diagonal. In order to correct for this, we need to subtract the sum of that diagonal, which is just n*1 (since each element in that diagonal equals 1):

Sum of the Correlation Matrix:

$$2\left(\frac{n(1-r)\ -\ r(1-r^n)}{(1-r)^2}\right)\ -\ n$$

This value is proportional to the sum of the covariance matrix, which is proportional to the variance of talent in the population over n days.

Next, we need to come up with a corresponding value to represent the variance of talent over a single day. To do this, we can rely on the fact that as long as talent never changes, the variance in talent over any number of days is the same as the variance in talent over a single day. Instead of comparing to the variance in talent over a single day, we can instead compare to the variance in talent over n days when talent is constant from day to day.

This allows us to construct a similar correlation matrix to represent the constant-talent scenario. Compared to the correlation matrix for changing talent, this is trivially simple: since talent levels are the same throughout the sample, the correlation between talent from one day to the next will always be one.

In other words, the correlation matrix will just be an n x n array of 1s. And the sum of an n x n array of 1s is just n^2.

Sum of the Correlation Matrix, changing talent:

$$2\left(\frac{n(1-r)\ -\ r(1-r^n)}{(1-r)^2}\right)\ -\ n$$

Sum of the Correlation Matrix, constant talent:

$$n^2$$

The ratio of these two values will give us the ratio of talent variance after n days of talent changes to the talent variance when talent is constant:

Reduction in variance of true talent over n days

$$\frac{var_{t_{0:n-1}}}{var_{t_0}} = \frac{2\left(\frac{n(1-r) - r(1-r^n)}{(1-r)^2}\right) - n}{n^2}$$

And that is our formula for finding the ratio of variance in true talent over n days to the variance in true talent on a single day, given the value r for the correlation of true talent from one day to the next. With some simplification, the above formula is equivalent to what was posted in the THT article:

Reduction in variance of true talent over n days

n ; number of days in the sample
r ; day-to-day correlation for true talent

$\boldsymbol{var_{t_{0:n-1}}}$; variance in talent over n days
$\boldsymbol{var_{t_0}}$; variance in talent over one day

$$\frac{\boldsymbol{var_{t_{0:n-1}}}}{\boldsymbol{var_{t_0}}} = \frac{\boldsymbol{n(1-r^2) - 2r(1-r^n)}}{\boldsymbol{n^2(1-r)^2}}$$