Population genetic models

Peter Beerli

October 17, 2011

To understand inferences based on sampling random relationships among a small sample of individuals of a contemporary population we need to know some basic models of population history. Several such models exist. Fisher (1929, and 1930) and Wright (1931) developed independently a simple population model. We call this model the *Wright-Fisher* population model. Alternatives are the Moran model, developed by Moran in 1958. A model that extends the Wright-Fisher model was described by Cannings in 1974. We will discuss the three models

Wright-Fisher population model

We assume that we have a population of N diploid individuals, so each individual has two gene copies at a specific locus. A population therefore has $2N$ gene copies. Every generation the individuals reproduce once and die. Each individual is releasing a very large number of gametes. The next generation is formed by picking random copies from this pool of gametes. Most simple implementation of this model do not allow for change of population size through time, mutation, selection, immigration and other complicating population genetics forces. If we assume that we start such a population with two alleles A_1 and A_2, than we look at the number of one specific allele X, say the number of A_1 alleles. The future states of X can represent any of $0, 1, 2, ..., 2N$. Sampling from the gene pool can be replaced by a sampling with replacement from the population at time t. This means that $X(t + 1)$ is a binomial random variable with index $2N$ and parameter $X(t)/(2N)$. We can express the transition probability from $X(t) = i$ to $X(t + 1) = j$ as

$$p_{ij} = \binom{2N}{j}(\frac{i}{2N})^j(1-(\frac{i}{2N}))^{2N-j}. \tag{1}$$

For example, in Figure **??** $2N = 10$, and we start with 5 individual chromosomes A_1 and 5 A_2, in the 5th generation we have 4 A_1 and 6 A_2 alleles. The probability to change from 4 A_1 (denoted

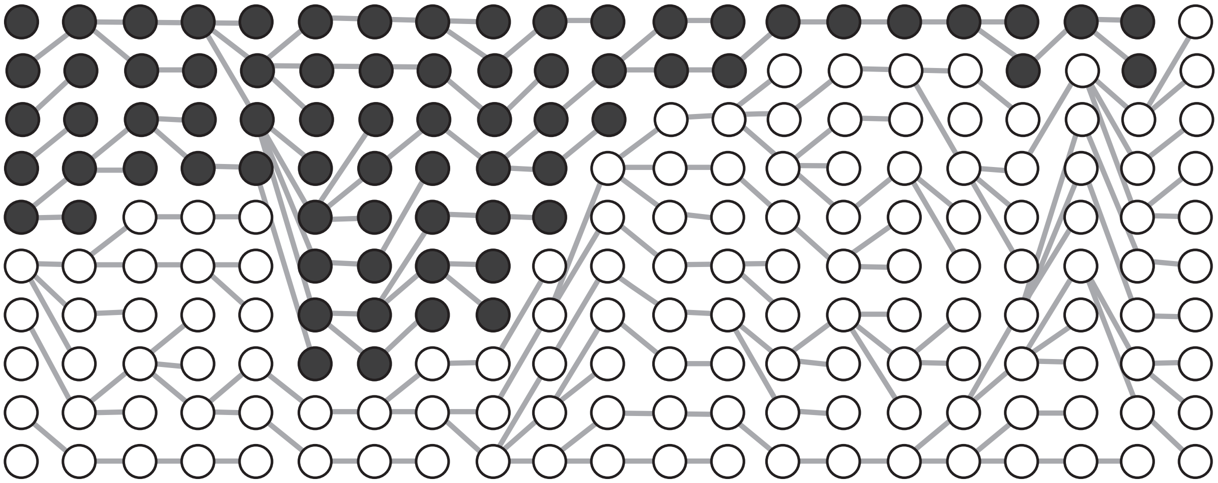

Figure 1: Example of a Wright-Fisher population model with two alleles, the past is on the left side, the gray lines denote ancestry.

as i in formula **??**) in generation 5 to 8 alleles (denoted as j) in generation 6 is

$$p_{4,8} = \binom{10}{8}(\frac{4}{10})^8(1-(\frac{4}{10}))^{10-8} = 0.0106168.$$

Interesting questions in this context are: (1) What is the chance that a specific allele fixes over time when it has frequency $i/(2N)$? (2) How long does it take to fix an allele? (3) How long does it take to fix a particular allele? (1) The chance of fixation is equivalent to the frequency the allele has at a particular time, for example the A_1 in the 5th generation in Figure **??** has frequency 4/10, and in 4 out of 10 trials A_1 will fix, Ewens (2004) gives 3 different way to arrive at that results, the simplest one suggest that when an allele just arises because of its low frequency it will be picked less often, and because of the finite number of draws the chance of getting lost is large, with intermediate frequencies small deviations from the expected number of picks do not matter, and with high frequency any number of drawing that are higher to the expectation increase the chance of fixation.

(2) The Wright-Fisher population model as explained here is prognostic, because we look into the future and this makes it particularly difficult to derive quantities as average to fixation, Ewens (2004) gives and example of the time of fixation $\bar{t}\,(p)$ of an allele with frequency p as

$$\bar{t}\,(p) = -4N\left(p\log p + (1-p)\log(1-p)\right), \tag{2}$$

when we insert a low frequency, one that arises if the population consists only of a single allele A_2 and a single mutation arises, A_1 then the time to fixation is

$$\bar{t}\left(\frac{1}{2N}\right) = 2 + \log 2N. \tag{3}$$

With equal frequency at the start generation the time to fixation is

$$\bar{t}\left(\frac{1}{2}\right) \simeq 2.8N. \tag{4}$$

If we assume that we are only interested in when allele A_1 fixes, then

$$\overline{t^*}\left(p\right) = -4N\frac{1}{p}\left(1-p\right)\log\left(1-p\right) \tag{5}$$

$$\overline{t^*}\left(\frac{1}{2N}\right) \simeq 4N - 2 \tag{6}$$

$$\overline{t^*}\left(\frac{1}{2}\right) \simeq 2.8N \tag{7}$$

$$\overline{t^*}\left(1-\frac{1}{2N}\right) \simeq 2 + \log 2N \tag{8}$$

Formula **??** is very different from formula **??**, of course because (**??**) only looks at half of the choices, fixation of allele A_1, the waiting time for any event fixation of A_1 or A_2 has to be shorter. The conditional and unconditional results are the same when the start allele frequency is 1/2.

Extending from two alleles to k alleles is easy as it uses simply the multinomial instead of the binomial distribution.

$$p(\{i_1, i_2, ..., i_k\}, \{j_1, j_2, ..., j_k\}) = \frac{(2N)!}{j_1!j_2!...j_k!}\left(\frac{i_1}{2N}\right)^{x_1}\left(\frac{i_2}{2N}\right)^{x_2}...\left(\frac{i_k}{2N}\right)^{x_k}, \tag{9}$$

$$\text{with } 2N = \sum_z^k j_z = \sum_z^k j_i \tag{10}$$

We shall later see that we can derive similar estimates using *coalescence theory*.

The transition probabilities $p(ij)$ allow a further insight into the behavior of these models. All possible states of (i, j) can be expressed in a transition matrix P. Using this matrix we can calculate the probability to move from state i, for example having 4 A_1 alleles and 2 A_2 alleles in a population of size $2N = 6$, to j, for example 3 A_1 and 3 A_2. In one generation this is 0.219. Waitin 2 generation to get the same transition has a probability of 0.173 and after 10 generation it is only 0.039. We can use the eigenvalues of this transition matrix the get some idea about how fast alleles are fixed (see description of eigenvalues at the end of this chapter). For the Wright-Fisher model the first non-trivial eigenvalue is

$$\lambda_2 = 1 - \frac{1}{2N} \tag{11}$$

The first two eigenvalues λ_0, λ_1 are 1.0 and indicate the two absorption states 0 and $2N$. The eigenvalue of the system describes the general loss rate but the two different calculation of time of fixation (unconditional and conditional) suggest that the eigenvalue itself is only a rough descriptor of the fixation process.

Detailed progress and more explanations can be found at the end of this chapter.

Canning's (Exchangeable) population model

The Canning's model can be viewed as an intermediate between the Wright-Fisher model and the Moran model. Because depending on the reproduction function it can mimic the other two models. In its most basic version it looks like the Wright-Fisher model . Consider a a "population" of genes of size $2N$ reproducing at time points $t_1, t_2, ...$ The transition between the old generation and the new generation can be very general (not like the Wright-Fisher model) as long as the model guarantees that all alleles at time t have the same distribution of descendants at time $t+1$: they need to have the same offspring probability distribution. This distribution has a mean of 1 offspring with variance σ^2. We can construct offspring distributions that fit this general description that are far from the multinomial distribution needed for the Wright-Fisher model.

The first non-trivial eigenvalue is

$$\lambda_2 = 1 - \frac{\sigma^2}{2N-1} \tag{12}$$

The time to fixation in the Canning's model is very similar to the Wright-Fisher model except that the time is scaled by the variance of the number of offsprings σ^2:

$$\bar{t}\,(p) = -(4N-2)\frac{(p\log p + (1-p)\log(1-p))}{\sigma^2} \tag{13}$$

$$\tag{14}$$

This makes immediate sense, for example when the variance is very large (much larger than 1) then the time to fixation shrinks because there is an increased chance compared to the Wright-Fisher model that one individual is the parent of all offspring, exterminating the other allele immediately. If the variance is smaller than 1 then the fixation time is longer because more parents will have offspring than under the Wright-Fisher model every generation, therefore delaying the loss of an allele

Moran's population model

Moran's model was derived for haploid populations, but many of its findings can be applied to diploids because if we assume neutrality then the alleles in a diploid population of size N behave like a haploid population of size $2N$. At time points $t_1, t_2, t_3, t_4, ...$, an "adult" individual is chosen at random to reproduce, after reproduction and "adult" individual is chosen randomly to die. Again with two alleles, A_1 and A_2, we can calculate transition probabilities. For comparison with the diploid Wright-Fisher model we still use $2N$ as the number of chromosomes in the population. If we have at time t i copies of allele A_1 then it at time $t+1$ there will be $i-1$ A_1 copies if an A_2 individual reproduced and an A_1 is chosen to die. This results in the probability

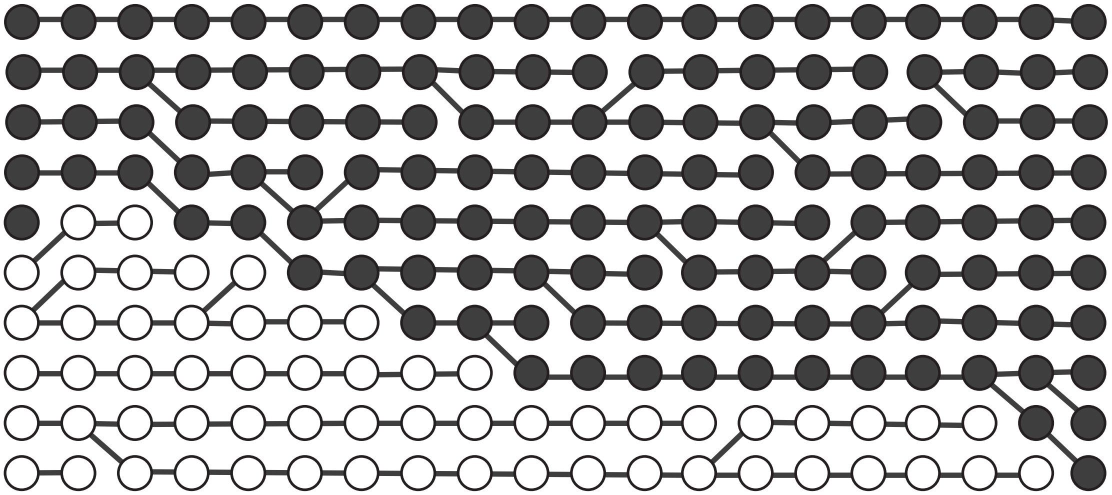

Figure 2: Example of Moran's population model with two alleles

$$p(i, i-1) = \frac{i}{2N}\frac{2N-i}{2N}. \tag{15}$$

With probability $i/2N$ an A_1-individual dies and it gets replaced by an A_2-indvidual with probability $(2N-i)/2N$. For gaining an A_1 (equals loosing an A_2 first) we find

$$p(i, i+1) = \frac{i}{2N}\frac{2N-i}{2N}, \tag{16}$$

and for no change

$$p(i, i) = \frac{i^2 + (2N-i)^2}{(2N)^2}. \tag{17}$$

The first non-trivial eigenvalue using the transition probabilities for the Moran model is

$$\lambda_2 = 1 - 2/(2N)^2 \tag{18}$$

The time to fixation of any allele of the Moran model with i A_1 and $2N-i$ A_2 alleles are

$$\overline{t_i} = 2N(2N-i)\sum_{j=1}^{i}\frac{1}{2N-j} + 2Ni\sum_{j=i+1}^{2N-1}\frac{1}{j} \tag{19}$$

but when we look only at the time of fixation for allele A_1: with i A_1 and $2N-i$ A_2 alleles, then we get

$$\overline{t_i^*} = 2N(2N-i)/i\sum_{j=1}^{i}\frac{j}{2N-j} + 2N(2N-i-1) \tag{20}$$

with 1 A_1 and $2N-1$ A_2 alleles

$$\overline{t_1^*} = 2N(2N-1) \tag{21}$$

Importance of time scale for comparison with the Wright-Fisher model

For a comparison with the Wright-Fisher population model we need to find a common measure of time as the events happening at times t_j in a Moran model are not equivalent to the events that happen at the times t'_j in a Wright-Fisher Model. We can express a common time g as the complete turnover of a population. If we set g=1 for the Wright-Fisher model then the Moran model needs on average around $2N$ time events to turn over, so we can express it generation time $g = 2N$. So we might express the waiting time to fixation in Wright-Fisher units as $\overline{t}_g(1/(2N)) = 2N-1$.

Comparison of the Wright-Fisher, Canning's, and Moran model

The Wright-Fisher model is widely used and most commonly taught, for analytical treatment it is rather difficult and often needs diffusion equation to solve simple problems, the Moran model is much more tractable and has many analytical solution that are difficult to come by with the Wright-Fisher or Canning model, but the Moran model is an extreme case of overlapping generation models. The Canning model allows for other variances of offspring number than the Wright-Fisher model that is fixed at

$$\sigma^2_{WF} \simeq 1 \tag{22}$$

and the Moran model that is fixed at

$$\sigma^2_{M} \simeq 2/(2N) \tag{23}$$

These different variances are responsible for the differences among the models, but most often simply function as a scaling quantity: in the Moran model the fixation time is about twice as fast as in the Wright-Fisher model, and the fixation time in the Canning model depends on the offspring variance σ^2, but is a simple scaling of the Wright-Fisher model by $1/\sigma^2$.

About transition probability matrices and eigenvalues

What are the chances to run many generations when the population size N is small, here $N = 3$, therefore we have 7 potential states with two alleles A_1, and A_2. The states of the system are $(0,6),(1,5),(2,4),(3,3),(4,2),(5,1)$, and $(6,0)$. We can calculate a transition probability matrix (see Generation 1), that calculates all transition probabilities using formula **??**

$$p_{ij} = \binom{2N}{j}(\frac{i}{2N})^j(1-(\frac{i}{2N}))^{2N-j}.$$

For example, in the first generation when we started out with 2 A_1 alleles and 4 A_2 alleles (2,4) and we want to know what is the probability after one generation to end up with 4 A_1 and 2 A_2 alleles $(4,2)$, that is 0.082. Now what is the chance to have start out with $(4,2)$ and after two generation being still at $(4,2)$? 0.108. Etc. The chances to stay in the same state gets smaller (Generation 1 to 3 : $(4,2) \rightarrow (4.2)$ is 0.036) and smaller every generation (Generation 1 to 4 : $(4,2) \rightarrow (4.2)$ is 0.002) because we loose to the absorbing states (the values in the matrix are only given up to 3 digits, and truncated to 0 when smaller than 0.001)

$$\text{Generation 1:} \quad \begin{pmatrix} 1. & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.335 & 0.402 & 0.201 & 0.054 & 0.008 & 0.001 & 0 \\ 0.088 & 0.263 & 0.329 & 0.219 & 0.082 & 0.016 & 0.001 \\ 0.016 & 0.094 & 0.234 & 0.312 & 0.234 & 0.094 & 0.016 \\ 0.001 & 0.016 & 0.082 & 0.219 & 0.329 & 0.263 & 0.088 \\ 0 & 0.001 & 0.008 & 0.054 & 0.201 & 0.402 & 0.335 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1. \end{pmatrix} \tag{24}$$

$$\text{Generation 2:} \quad \begin{pmatrix} 1. & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.488 & 0.22 & 0.16 & 0.084 & 0.035 & 0.011 & 0.002 \\ 0.208 & 0.214 & 0.22 & 0.174 & 0.111 & 0.054 & 0.018 \\ 0.073 & 0.133 & 0.189 & 0.211 & 0.189 & 0.133 & 0.073 \\ 0.018 & 0.054 & 0.111 & 0.174 & 0.22 & 0.214 & 0.208 \\ 0.002 & 0.011 & 0.035 & 0.084 & 0.16 & 0.22 & 0.488 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1. \end{pmatrix} \tag{25}$$

$$\text{Generation 3:} \begin{pmatrix} 1. & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.577 & 0.139 & 0.12 & 0.082 & 0.048 & 0.024 & 0.01 \\ 0.302 & 0.162 & 0.166 & 0.141 & 0.108 & 0.071 & 0.049 \\ 0.137 & 0.126 & 0.155 & 0.163 & 0.155 & 0.126 & 0.137 \\ 0.049 & 0.071 & 0.108 & 0.141 & 0.166 & 0.162 & 0.302 \\ 0.01 & 0.024 & 0.048 & 0.082 & 0.12 & 0.139 & 0.577 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1. \end{pmatrix} \tag{26}$$

$$\text{Generation 4:} \begin{pmatrix} 1. & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.635 & 0.096 & 0.091 & 0.071 & 0.051 & 0.032 & 0.024 \\ 0.374 & 0.124 & 0.13 & 0.117 & 0.098 & 0.073 & 0.085 \\ 0.196 & 0.109 & 0.128 & 0.133 & 0.128 & 0.109 & 0.196 \\ 0.085 & 0.073 & 0.098 & 0.117 & 0.13 & 0.124 & 0.374 \\ 0.024 & 0.032 & 0.051 & 0.071 & 0.091 & 0.096 & 0.635 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1. \end{pmatrix} \tag{27}$$

$$\text{Generation 5:} \begin{pmatrix} 1. & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.676 & 0.07 & 0.07 & 0.06 & 0.048 & 0.035 & 0.041 \\ 0.429 & 0.097 & 0.104 & 0.097 & 0.086 & 0.068 & 0.12 \\ 0.246 & 0.092 & 0.107 & 0.109 & 0.107 & 0.092 & 0.246 \\ 0.12 & 0.068 & 0.086 & 0.097 & 0.104 & 0.097 & 0.429 \\ 0.041 & 0.035 & 0.048 & 0.06 & 0.07 & 0.07 & 0.676 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1. \end{pmatrix} \tag{28}$$

$$\text{Generation 6:} \begin{pmatrix} 1. & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.707 & 0.053 & 0.055 & 0.05 & 0.043 & 0.033 & 0.058 \\ 0.472 & 0.077 & 0.084 & 0.081 & 0.074 & 0.061 & 0.152 \\ 0.288 & 0.077 & 0.089 & 0.091 & 0.089 & 0.077 & 0.288 \\ 0.152 & 0.061 & 0.074 & 0.081 & 0.084 & 0.077 & 0.472 \\ 0.058 & 0.033 & 0.043 & 0.05 & 0.055 & 0.053 & 0.707 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1. \end{pmatrix} \tag{29}$$

$$\text{Generation 7:} \begin{pmatrix} 1. & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.749 & 0.033 & 0.036 & 0.035 & 0.032 & 0.027 & 0.088 \\ 0.534 & 0.05 & 0.056 & 0.056 & 0.053 & 0.045 & 0.205 \\ 0.353 & 0.054 & 0.062 & 0.063 & 0.062 & 0.054 & 0.353 \\ 0.205 & 0.045 & 0.053 & 0.056 & 0.056 & 0.05 & 0.534 \\ 0.088 & 0.027 & 0.032 & 0.035 & 0.036 & 0.033 & 0.749 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1. \end{pmatrix} \tag{30}$$

$$\text{Generation 8:}\quad \begin{pmatrix} 1. & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.749 & 0.033 & 0.036 & 0.035 & 0.032 & 0.027 & 0.088 \\ 0.534 & 0.05 & 0.056 & 0.056 & 0.053 & 0.045 & 0.205 \\ 0.353 & 0.054 & 0.062 & 0.063 & 0.062 & 0.054 & 0.353 \\ 0.205 & 0.045 & 0.053 & 0.056 & 0.056 & 0.05 & 0.534 \\ 0.088 & 0.027 & 0.032 & 0.035 & 0.036 & 0.033 & 0.749 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1. \end{pmatrix} \tag{31}$$

$$\text{Generation 9:}\quad \begin{pmatrix} 1. & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.764 & 0.027 & 0.03 & 0.029 & 0.028 & 0.023 & 0.1 \\ 0.557 & 0.041 & 0.047 & 0.047 & 0.045 & 0.039 & 0.226 \\ 0.377 & 0.045 & 0.051 & 0.053 & 0.051 & 0.045 & 0.377 \\ 0.226 & 0.039 & 0.045 & 0.047 & 0.047 & 0.041 & 0.557 \\ 0.1 & 0.023 & 0.028 & 0.029 & 0.03 & 0.027 & 0.764 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1. \end{pmatrix} \tag{32}$$

$$\text{Generation 10:}\quad \begin{pmatrix} 1. & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.776 & 0.022 & 0.024 & 0.024 & 0.023 & 0.02 & 0.111 \\ 0.575 & 0.034 & 0.039 & 0.039 & 0.038 & 0.032 & 0.243 \\ 0.398 & 0.037 & 0.043 & 0.044 & 0.043 & 0.037 & 0.398 \\ 0.243 & 0.032 & 0.038 & 0.039 & 0.039 & 0.034 & 0.575 \\ 0.111 & 0.02 & 0.023 & 0.024 & 0.024 & 0.022 & 0.776 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1. \end{pmatrix} \tag{33}$$

The change of the matrix can be explored differently, by calculating the eigenvalues and eigenvector of the matrix. We make a matrix transform so that we split the matrix into 3 matrices of articular form

$$P = Z^{-1}\Lambda Z \tag{34}$$

You should think of this like a transformation of the coordinate system where the Z matrix now contains the new axes and the Λ contains the scalers of these axes; Λ is a matrix with eigenvalues on the diagonal, for our example it looks like this

$$\begin{pmatrix} 1. & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1. & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.833 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.556 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.278 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.093 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.015 \end{pmatrix} \tag{35}$$

The ordering of eigenvalues (and eigenvectors) often goes from the largest eigenvalue to the smallest. Here, the first two entries with 1 are for the absorbing states (0 and $2N = 6$), the next one is 0.8333. In the Wright-Fisher population model we have a first non-trivial eigenvalue of $1 - 1/(2N)$ and with $N = 3$ this results is exactly the value we calculated: 0.833.