An Introduction to Evolutionary Biology

Prof. Sutirth Dey

Biology Department, Population Biology Lab

Indian Institute of Science Education and Research (IISER) Pune

Week 3 Lecture 14

Solving simple Hardy-Weinberg problems

So, in our last discussion, we talked about why the Hardy-Weinberg equilibrium is so important in the context of evolutionary biology. And I explicitly said that in this discussion. We are going to look at some simple numerical problems using the Hardy-Weinberg principle and how to solve them. Now, why exactly do we want to do that? Obviously, you know, one reason is that it is a very important concept. The second reason is that, from a very pragmatic point of view, Hardy-Weinberg questions Numerical questions very often feature in national-level exams like NET or GATE. Or, you know, even in international exams like the TOEFL—sorry, not TOEFL, GRE, etc—typically, they feature them. Even most interviews, whether in the field of evolution or in the field of genetics, Many a time, people do end up asking Hardy-Weinberg questions. Now, solving the Hardy-Weinberg numerical problems is actually very simple. However, it is my experience that students make some very trivial kinds of mistakes because of which often things you know do not remain as simple as they are supposed to be. So, what I will do is we will first look at something simple; you know, we will basically play with the Hardy-Weinberg frequencies a bit. so that you get a feel for what those frequencies, you know, look like. And then I am going to do three or four problems, each one of them a slightly different type. And I will not only show you how to solve those kinds of problems, but I will also quickly talk about what the common mistakes are.

The sentence is incomplete and does not provide enough context for a correction. So, the first thing is looking at, you know, looking at the Hardy-Weinberg frequencies. So, for this, what we will do is go to Excel. Now, I am doing it on Microsoft Excel; you can also do it on the corresponding LibreOffice. Any other office software, more or less, will operate the same way.

So, what do we want to do? What we want to do is take various allele frequencies, and for those, we will plot the corresponding gene frequencies. So, allele frequencies let us say we start with 0. I will slightly increase the size so that you can see it better. So, we will start with 0 and then we will increase in steps of 0.01. So, 0, 0.01, 0.02, and then we are just going to drag this all the way up to 1. Now, remember this is a fraction in the sense that this is our frequency. Therefore, the values it can take can only be between 0 and 1.

So, this is p. Now, we are going to look at q, which is (1-p). So, we will simply say this is 1 minus this. And then we are also going to drag it. Okay. So now we have our allele frequencies. Now, we want to calculate our genotypic frequencies. And remember, the genotypic frequencies are p^2, 2pq, and q^2. So, we will simply type [= p*p], [= 2*p*q], and [= q*q], right? And now let's simply drag this all the way to the end. And there you go.

Okay. So, now what we would like to do is plot the three genotypic frequencies against the corresponding allele frequencies. So, how do we do that? In order to do that, we go to insert, and we insert this kind of graph. And then we go to select data, and then we look at the chart data range. Just one second, we need to shift this a bit towards this side. So, we'll go to chart design and select the data range.

So, the data range is the entire range of data from here to here. Done. And then on the x-axis, sorry, this is the vertical axis. So, on the vertical axis, we need p^2, 2pq, and q^2. And on this horizontal axis, we are going to say that we need from here to here.

Okay. And that's it. So, if you remember, this is the graph that we looked at yesterday when we were, I am sorry, I mean in the previous discussion. When we were looking at how the genotypic frequencies vary with the allele frequencies, right? So, of course, one can play with it a bit; one can change the, you know, axis instead of going from 0 to 1.2. One can make it go from 0 to 1, and so on and so forth. Sorry. Yeah, 0 to 1, and this goes to 0. Right, there you go. So, why am I showing you this? There are two reasons for this. Number one, of course, it is very nice to actually know, you know, about the plots that you see during lectures. It is nice to be able to draw them on your own.

So, I strongly recommend that you do this. I am doing this in Excel. If you know how to write a program, please feel free to write Python code or any other language you are familiar with. But draw this; this is fun. The other reason is that normally students only think in terms of p^2, 2pq, and q^2 without realizing that These, actually, you know, stand for numbers; these are algebraic expressions that stand for numbers.

And some of these numbers, if you actually keep them in mind, become very easy for you to solve Hardy-Weinberg numericals. So, for example, look at this one—the one that I am highlighting over here. Yeah, so you can see that when the allele frequencies are 0.1 and, therefore, 0.9 for the other one, Then you get nice whole numbers: 1%, 18%, and 81%.

Similarly, when you know they are 0.2 and 0.8, you again get nice values: 0.04, 0.32, and 0.64. So, these are the ones that stand for, you know, 0.1, 0.2, 0.3, 0.4; these are the ones that are very often favored by question setters, simply because taking the square root of these things becomes very, very easy. And in all those kinds of exams where calculators are not made available to students, you know, These kinds of questions are the ones that allow people to give a Hardy-Weinberg equilibrium with understanding. That you know if it is 0.49 or you know something, then it will be relatively easier for people to take the square root of that. You know, the square root of 0.49 is 0.7 and so on. So, these values that I am highlighting, I highly recommend you keep these values in mind, okay. So, once you see any of these genotypic values, You automatically know what the allele frequencies are; you know, without even thinking. This does not mean that I want you to mug up all the values; no, I just want you to mug up these three or four that I highlighted.

Okay, so with this background, we can now afford to shift to the problems. Okay, so here is the first problem.

Q1. The following numbers were noted while studying the MN blood group in a population: 83 MM, 46 MN, and 11 NN. Compute the gene and genotypic frequencies.

Solution: So we are talking about 83 + 46 + 11. So, 83 + 11 is 94, and 94 + 46 = 140. So this total is 140 individuals, right? Now, a very common mistake that students make is taking 140 individuals, then calculating 11/140 of something. They equate this to q^2, and then they tend to take the square root of this; thereby, they derive the gene frequencies. Now, this is actually wrong. Why is it wrong? Because this q^2 representing the genotypic frequency presupposes that the population is in Hardy-Weinberg equilibrium. Whereas in this particular case, note that they haven't really told you that the population is in Hardy-Weinberg equilibrium, right? So, the basic thumb rule here is that any time you have access to all the genotypic numbers For all the genotypic frequencies, you should never use the Hardy-Weinberg p^2, $2pq$, q^2 formula. What you should instead use is the small $p = (P + Q / 2)$ formula because this relationship, remember the way we derived it, This relationship actually does not require the Hardy-Weinberg equilibrium to be reached in the population, right? So, how are you going to do this? So, in this context, we will first compute the genotypic frequencies. So, the genotypic frequency for MM is simply 83/140, and that for MN is 46/140, okay? So, this is 83, this is 46, and similarly for NN, the genotypic frequency is 11/140. So, I happen to have a calculator near me. So, this is my calculator. Let us quickly do this. So, 83/140 = 0.592. Similarly, 46/140 = 0.329, and 11/140 = 0.079, actually 0.0785, but I will just round it to 0.079. So, these are my genotypic frequencies. And from these genotypic frequencies, I am going to use this formula to get my p and q values.

So, $p = P + (1/2) * Q$; $p = 0.592 + (1/2) * 0.329$, let us see. So, 0.329 * 0.5 is 0.1645, and 0.1645 + 0.592 = 0.7565. Therefore, I will represent that as 0.757.

And if p is 0.5757, then q is equal to (1-0.757). So, (1-0.757) is equal to 0.243.

And if you really want to check this, you can also get to roughly the same value by using 0.079 for q: q = 0.079 + (1/2) * 0.329. So, let us see what that leads to. So, (0.5*0.329)=0.1645, (0.1645 + 0.079)=0.2435.

I mean, this is just, you know, this is pretty close to this. Whatever you are seeing over here in the fourth digit is just a rounding off error. So, this tells you that as long as you have all the genotypes, make sure that you use this formula to go from genotypic frequency to gene frequency. Okay, next one. So, if you remember, we said that this Hardy-Weinberg principle that we derived for the 1-locus 2-allele case, It actually extends very nicely to the 1-locus multi-allelic case.

And I told you that the various genotypic coefficients for the frequencies will be, you know, $(p+q+...+n)^2$, the coefficients of that expansion. So, just to give you an example, actually in this particular case, we do not even need to go all the way to the, you know, expansion. We can do it much more simply, but the principle will be the same as we saw last time.

Q2. So, the question is this: There are three allelic variants, A, B, and C, of the red cell acid phosphatase enzyme found in a sample of 178 English people.

Solution: This number is important, we will come back to it. All genotypes were distinguishable by electrophoresis, and the frequencies in the sample are as follows. So, they have given you the frequencies, but note that they have given them in percentages, right? So, in order to do the calculation, we will have to convert this into the corresponding 0-to-1 range.

How does one do that? Essentially, dividing everything by 100, right? So, what are the values? This is going to be 0.096, this is going to be 0.483, this is going to be 0.343. This is going to be 0.028, and this is going to be 0.05, and this will remain as 0, right? So, these are the frequencies. Now, from these, we need to know the frequencies of the three alleles A, B, and C in the sample. So, we have to go from genotype frequency to allele frequency. How do we get there? Remember, all the genotypic frequencies are given here. So, you do not really need to assume the square root kind of relationship. You can simply go there by our p = P + Q/2, except that now instead of Q/2. You are going to do it as Q plus, whatever, you know, Q plus Q1 or whatever. Basically, these are the heterozygotes which also contain the particular allele in question. In this particular case, let us say for allele A, these are going to be the heterozygotes containing allele A, which are AB and AC. So, just to show you how it works: let p be the frequency of allele A, and we will call it pA to make it really easy.

Frequency of A = [frequency of the homozygote, which is 0.096, + 1/2 * the frequency of the two heterozygotes that contain A.] which are AB (0.483) and AC (0.028)], which basically means 1/2*(0.483 + 0.028), right? So, how much does that boil down to? Let us do this again. So, (0.483 + 0.028), my calculator tells me it is 0.511, and we multiply that by (1/2), which gives me 0.2555. Then to this gets added, (0.2555 + 0.096) = 0.3515. So, this is equal to pA.

Similarly, pB = [the frequency of the homozygote, BB homozygote, which is 0.343 + 1/2 * the frequency of the two heterozygotes] which contain B which happen to be BC and AB which means (0.483 for AB + 0.05 for BC) =? Let us see what it comes to. So, (0.483 + 0.05) = 0.533, (0.533 * 0.5) = 0.2665, (0.2665 + 0.343) = 0.6095. So, pB = 0.6095.

And now we come to pC: pC = [0 + 1/2 * (which are the two heterozygotes AC and BC)], right? The ones that contain C, (0.028 + 0.0505). So, let us look at the calculator. So, (0.028 + 0.05) = 0.078, (0.078 * 0.5) = 0.039, and 0.039 + 0 = ? So, basically, this comes down to 0.039.

Now, whenever you do calculations like this, it is always a good idea to add these numbers up. Because if I have 3 alleles, these 3 numbers should add up to 1. So, 0.3515 + 0.6095 + 0.039; sorry, I made a mistake over here. (0.3515 + 0.6095 + 0.039) = 1; as you can see, this is equal to 1, right? So, these calculations are fine. That is part A. Then this asks the question, why are no CC individuals found? Now, in order to answer this question, we have to ask ourselves what the expected frequency of the CC individuals is. So, my allele frequency of the C allele is 0.039. Therefore, my expected genotypic frequency of C is (0.039*0.039) = 0.001521, right? This is my expected frequency of the CC homozygotes. Now remember, I have only 178 individuals in my sample. So in that sample, how many CC individuals do I expect? In order to get to that figure, we multiply 0.00152 by 178 to get 0.27. In other words, I expect only 0.27 individuals in a sample of 178. Since this number is less than 1, it essentially means that I do not even expect a CC individual in the sample. So, this example shows you that the sample size is also very important. When you are trying to get your gene frequencies and genotypic frequencies from real-life data. If a gene is really rare or an allele is really rare, then in order to see their corresponding homozygotes, you need to have a very large sample size; otherwise, the homozygote typically will not be seen. So this is how you extend this to a one-locus multi-allelic case. So in both these examples, all the genotypic frequencies are given. Which is why we did not really need to use the Hardy-Weinberg principle.

Q3. Now we will look at a case where all the genotypic frequencies are not given. So, as the question states, a 1972 study shows that the frequency of phenylketonuria patients in a population is 1 in 11,000. Disease is caused by an autosomal recessive gene. What is the frequency of A. the disease gene, and B. the frequency of the carriers, that is, the heterozygotes?

Solution: So in this particular case, note that you have been given only one genotypic frequency. Therefore, you do not have all the other genotypic frequencies in hand. Hence, you cannot use your $p = P + Q / 2$ formula; therefore, in this particular case, The only way for you to proceed is to explicitly assume that the population is in Hardy-Weinberg equilibrium. And then take the square root to obtain the gene frequency of the disease gene. Now in many cases, I have seen that students end up doing this. Without explicitly stating that they are assuming the Hardy-Weinberg equilibrium. I would strongly recommend that when you are doing a descriptive type of question, you should mention this. You are, you know, assuming Hardy-Weinberg because otherwise, that taking of that square root is not justified. So anyway, in this particular case, let us assume that our two genes are A1 and A2, such that A2 is the disease gene.

Let us assume the frequency of this is p, let us assume the frequency of this is q, and therefore, as per the question, $q^2 = 1/11000$. So, of course, $q = \sqrt{(1/11000)}$. Let us quickly see what that number is. So 1/11000 is 0.0000909 and therefore when I take the square root of this, that comes out to be 0.0095. I mean there are more digits, but I am just cutting it off at 4 digits. So if this is my q, then my p = (1 - 0.0095) = 0.990, and then what I can do is my 2*p*q = 2 * 0.990 * 0.0095 = ? right? So, 2 * 0.990 * 0.0095 = 0.0189, roughly. So this is the frequency of the homozygote; this answer is 0.0189, and this answer is 0.0095. So the interesting thing to note in this particular question is that although the frequency of this disease and the diseased allele is very low. In spite of that, about 0.0189, which roughly means about 1 in 20 people, are actually going to be diseased. I am sorry, we are going to be carriers, which tells you that when a disease gene becomes rare. Then most of the copies of that rare gene are actually going to be found in the carriers. and that is why if you have a recessive gene it is very difficult to throw it out of the population by the action of selection alone. Selection won't even be able to see it. Okay, let's go to the next question.

Q4. So, the question states that in a given population, 35.42% of the individuals are carriers of a genetic condition. Assume that the disease is due to a single autosomal recessive gene and that the population is in Hardy-Weinberg equilibrium. If the recessive gene is rarer than the dominant gene, what is the expected percentage of individuals with that condition?

Solution: So, note what is happening in this one. Here they are giving you the percentage of the carriers that are the heterozygotes. They are asking you for the expected percentage of individuals with that condition. So, which basically means this is the inverse of the previous question. There they had given you the diseased condition, the homozygotes, and they had asked you about the frequency of the carriers. Here it is the other way around. So, it turns out that this is a very simple way of doing these questions. So, as per this question, 2pq is equal to 35.42%, which means 0.3542, right? Here they have already given you the Hardy Weinberg equilibrium, so you do not have to bother; it is right there. The question also says that the recessive gene is rarer than the dominant gene. It is not entirely clear right now where that will be important, but we will see that in a moment. So, 2pq = 0.3542. Now, let us assume that this is our rarer gene. So, this is, let us say, A2, and let us say this is the dominant gene. This is the dominant gene, and this is the recessive gene. So, we need the expected percentage of individuals with that condition.

So, basically, we need to go to q^2. In order to go to q^2, we first have to figure out what q is.

So, if that is the case, let us say that 2 * (p as 1 - q) * q = 0.3542 ||| 2q(1 - q) = 0.3542 So, we can divide both sides by 2, which will basically make it $(q - q^2) = 0.3542 / 2 = 0.1771$. So, taking everything to the right-hand side, $q^2 - q + 0.1771 = 0$. Now, we know that there is a very famous formula that says that if $ax^2 + bx + c = 0$, Then $x = [-b \pm \sqrt{(b^2 - 4ac)}]/2a$, and ± means that we are going to get two values from this where a, b, and c are the coefficients as written over here. So, applying that, the value of q = -b. So, my b over here is already -1, right? Because this is a minus sign over there. So, -(-b) = +b, which basically means $[1 \pm \sqrt{(b^2 - 4ac)}] / 2a$; So, b is 1. So, (1-4ac); a is again 1 here. So, (4*1*0.1771)/2. So, this stuff inside the brackets, this is (0.1771 * 4) = 0.7084. So, subtracting that from 1, (1 - 0.7084) = 0.2916, and then taking the square root of that gives me $\sqrt{0.2916} = 0.54$. So, this is boiling down to (1±0.54)/2, which is equal to dividing both numbers by 2. So, (0.5 ± 0.27), half of 0.54 is 0.27. If I add 0.27 to this, then I get two answers here. One is (0.5 + 0.27) = 0.77, and the other is (0.5 - 0.27) = 0.23. So, there are two possible answers for this value of q, which are 0.77 and 0.23.

And this is where this extra condition comes in. If the recessive gene is rarer than the dominant gene, it means that the frequency of the recessive gene is lower. Therefore, this is the correct answer given that condition. In other words, for this question, we are going to say that the expected percentage of individuals with that condition is $q^2 = (0.23)^2 = (0.23 * 0.23) = 0.0529$. Roughly speaking, $q^2 \sim 0.053$. So, this is in the frequency domain. Now, it states the expected percentage. So, that is going to be about 5.3 percent. This is the final answer to the question. Now, note here that if this particular condition—you know, this one—if the recessive gene is rarer than the dominant gene, This particular condition was not given, then you had no way of choosing between 0.77 and 0.23. And in that case, even for the expected fraction, you would have had two answers, okay. One is 5.3 percent; the other is $(0.77)^2$, and that is perfectly okay. If they do not give you an extra condition, you cannot distinguish between these two cases. So, that is about it. We are going to stop here, and in the next discussion, We are going to look at the ways in which genetic variation is generated and the implications thereof. See you. Bye.