Geographic Population Structure

Alan R. Rogers

June 15, 2024

In Buri's [1] drift experiment, heterozygosity (H) declined. At the same time, the variance (V) among populations increased. We already have a model describing the first of these phenomena. Here, we consider the second—the variance among populations—assuming throughout that genetic drift is the only evolutionary force at work.

First a few terms. In the initial generation, the allele frequency was p_0. We will treat this as a constant, not a random variable. In generation t, the corresponding quantity p_t is a random variable, the randomness having been introduced by the process of genetic drift. For convenience, we set $q_t = 1 - p_t$. The heterozygosity in generation t (also a random variable) is $H_t = 2p_tq_t$. We are interested in the evolutionary change that has happened since generation 0.

1 Genetic drift does not change the expected allele frequency

Let us begin with the expected value of p_1, the allele frequency in the generation 1. According to the urn model, the number of copies of allele A in each subpopulation is a Binomial random variable with mean Np_0. The expected allele frequency is thus $Np_0/N = p_0$. This demonstrates a remarkable fact: the expected allele frequency is unchanged by genetic drift. This is true not only of the first generation, but of each succeeding generation. No matter how many generations are involved, the allele frequency of each subpopulation is a random variable whose expected value is p_0. In symbols,

$$E[p_t] = p_0. \tag{1}$$

2 The Wahlund Principle and F_{ST}

In contrast to p_t, the variance V_t has an expected value that does change with time. To model the variance, we begin with its definition:

$$V_t = E[p_t^2] - p_0^2$$

This is just the standard definition of the variance, modified slightly in light of Eqn. 1. Note that, in view of this definition,

$$E[p_t^2] = V_t + p_0^2 \tag{2}$$

We'll need this fact in a minute.

Let us turn now to the heterozygosity. Its expected value in generation t is

$$\begin{aligned} E[2p_tq_t] &= 2E[p_t] - 2E[p_t^2] \\ &= 2p_0 - 2(V_t + p_0^2) \qquad \text{using Eqns. 1–2} \\ &= 2p_0q_0 - 2V_t \end{aligned} \tag{3}$$

This is another important fact: average heterozygosity in generation t will be smaller than that in generation 0. The amount of the reduction is *exactly twice* the variance of group allele frequencies about their expected value p_0. This is known as Wahlund's principle [3]. It shows that there is a close and necessary connection between the decline of heterozygosity (shown on the left in Fig. 1) and the increase in variance (shown on the right). In effect, genetic drift converts heterozygosity into variance among groups.

It is often useful to express the absolute reduction in heterozygosity ($2V_t$) as a proportion of the original heterozygosity ($2p_0q_0$). This proportional reduction is

$$F_{ST} = \frac{2V_t}{2p_0q_0} = \frac{V_t}{p_0q_0} \tag{4}$$

The notation F_{ST} was introduced by Sewall Wright [4] and is now conventional within population genetics. Gillespie defines F_{ST} using different notation

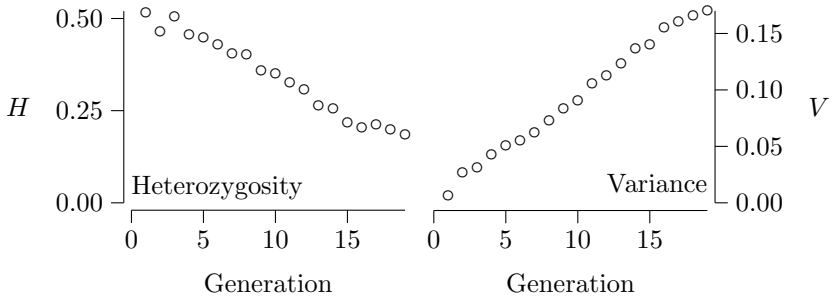

Figure 1: In Buri's [1] drift experiment, heterozygosity (H) declined. At the same time, the variance (V) among populations increased. Data are from Buri's series I and are tabulated in Table 1.

(his Eqn. 5.3), which was introduced by Nei [2]. Although the two definitions look different, they are interchangeable. In view of Eqn. 4, we can re-express Eqn. 3 as

$$E[2p_tq_t] = 2p_0q_0(1 - F_{ST}) \tag{5}$$

Now $E[2p_tq_t]$ is the expected heterozygosity in generation t, and $2p_0q_0$ is that in generation 0. Thus, Eqn. 5 says that F_{ST} is the proportional reduction in heterozygosity caused by genetic drift. There are corresponding increases in expected homozygosity. The expected frequencies of the three genotypes at a biallelic locus are:

Genotype	Frequency
AA	$E[p_t^2] = p_0^2 + p_0q_0F_{ST}$
Aa	$E[2p_tq_t] = 2p_0q_0(1 - F_{ST})$
aa	$E[q_t^2] = q_0^2 + p_0q_0F_{ST}$

These formulas are identical to those for pedigree inbreeding, because genetic drift is a form of inbreeding.

3 Model of completely isolated subpopulations

In Eqn. 5, we express the expected heterozygosity ($2p_tq_t$) in generation t as a function of that in generation 0. Early in the semester, we derived a similar result:

$$E[2p_tq_t] = 2p_0q_0\big(1 - 1/2N\big)^t \tag{6}$$

To refresh your memory, see Gillespie's Eqn. 2.3. These equations look very different, yet both are correct. They provide a simple way to derive the rule by which F_{ST} changes with time. Set Eqns. 5 and 6 equal to each other, and solve for F_{ST}. You will discover that

$$\begin{aligned} F_{ST} &= 1 - \big(1 - 1/2N\big)^t \\ &\approx 1 - e^{-t/2N} \end{aligned} \tag{7}$$

The last line above uses the approximation that $e^x \approx 1 + x$ if x is near zero. Eqn. 7 applies when the populations are totally separated. It shows that F_{ST} increases according to a very simple rule, increasing gradually toward its maximal value, 1.0.

4 Migration in addition to drift

If there is migration between populations, we cannot use any of the results in section 3. This case demands a different theory, which is discussed by Gillespie. Under the "island-model" of population structure, F_{ST} converges toward an equilibrium at which

$$F_{ST} = \frac{1}{4Nm + 1} \tag{8}$$

as shown on Gillespie's p. 136.

We have two equations for F_{ST}. Eqn. 8 refers to equilibrium under the island model, and Eqn. 7 to the case of totally isolated populations. In the exercises below, we will consider human data under these two extreme cases.

Exercises

1. Plot the V_t data in Table 1, to make a graph like that on the right side of Fig. 1.

2. Combine Eqns. 4 and 7 to obtain a formula for the variance, V_t. Plot this formula as a line in your graph, assuming as Buri did that $2N = 18$. How well do Buri's variance data fit the model?

3. In the continental human populations, $F_{ST} \approx 1/9$. Use this value to estimate Nm (under the equilibrium model).

Table 1: Data from series I of Buri's [1] drift experiment, as plotted in Fig. 1. Key: t, generation; H_t, mean heterozygosity ($2pq$) within subpopulations; V_t, variance among group allele frequencies.

t	H_t	V_t	t	H_t	V_t
0	1.000	0.000	10	0.348	0.090
1	0.514	0.006	11	0.325	0.105
2	0.464	0.026	12	0.305	0.112
3	0.504	0.031	13	0.263	0.123
4	0.456	0.042	14	0.255	0.136
5	0.448	0.050	15	0.216	0.140
6	0.428	0.055	16	0.202	0.155
7	0.403	0.062	17	0.210	0.160
8	0.402	0.072	18	0.197	0.165
9	0.358	0.083	19	0.183	0.170

4. Now assume that the human continental populations have been totally isolated, and use the observed F_{ST} to estimate $t/2N$. Then convert this into an estimate of the time in years since the human populations separated. (Assume that $N = 10,000$ and that generations are 25 years long.)

References

[1] P. Buri. Gene frequency in small populations of mutant *Drosophila*. *Evolution*, 10:367–402, 1956.

[2] Masatoshi Nei. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences, USA*, 70(12):3321–3323, 1973.

[3] Sten Wahlund. Composition of populations and of genotypic correlations from the viewpoint of population genetics. In Kenneth M. Weiss and Paul A. Ballanoff, editors, *Demographic Genetics*, pages 224–263. Wiley, New York, 1975. [Translation].

[4] Sewall Wright. The genetical structure of populations. *Annals of Eugenics*, 15:323–354, 1951.