Chapter 5

Random genetic-drift

Genetic drift is an evolutionary force that changes both allele and genotype frequencies. No population can escape its influence. John Gillespie

5.1 Offspring number is a random variable

So far we have been treating population and evolutionary dynamic as deterministic. That is, we assumed demographic and evolutionary processes without any variance. In reality, however, both of these processes will exhibit fluctuations due to various **stochastic** effects (or chance or random effects). Fundamentally, stochasticity is caused by the variance in the number of offspring produced by individuals. The fact that there is a variance in offspring number has a crucial consequence for evolutionary dynamics and leads to the phenomenon of **random genetic-drift**. This is the change in allele frequency (and thus evolution) even in the absence of natural selection and/or mutation.

In order to capture stochastic effects, we now let individual fitness be a random variable and denote it by the capital letter $W \in \{1, 2, ...\}$ the (random) number of offspring (possibly including self) produced by an individual. A **random variable** is a variable that can take on values from a set of possible values, each associated with a probability (see Appendix 1 for the probability concepts that will be used in this course). For instance, if the fitness of each individual follows an independent Poisson distribution (a common distribution for fecundity), then the probability that a particular individual produces k offspring; that is, the probability that the realized value of W is k is given by

$$\Pr(W = k) = \frac{\exp(-w)\,(w)^k}{k!}, \tag{5.1}$$

where

$$w = \mathrm{E}_{\mathcal{P}}[W] \tag{5.2}$$

is the expected fitness of the individual. The expectation $\mathrm{E}_{\mathcal{P}}[W]$ has subscript $\mathcal{P}$, which means that we condition on the parental generation. Indeed, we calculate the average fitness of an individual given the parental generation population state (e.g., the number of individuals, the genotypes and/or phenotypes of individuals, etc.)[1]. This is so because we always look at evolution over one demographic time step, from parent to offspring generation, and the information we have is the state of the parental generation. Hence, $\mathcal{P}$ can be thought to represent any information we have concerning the parental population, and every average we evaluate is conditional on that information.

Expected fitness is $w = \mathrm{E}_{\mathcal{P}}[W]$, which is exactly the fitness w we used previously (in chapters 2-4). But we have now made explicit that this expected fitness considers all stochastic events affecting reproduction and survival given parental population state $\mathcal{P}$. What we neglected in previous chapters, however, is that there is also a variance in the fitness of an individual, given the parental population state $\mathcal{P}$, and this is denoted $\mathrm{Var}_{\mathcal{P}}[W]$. Furthermore, there may also be a covariance in fitness between pairs of individuals. For the above Poisson distribution (eq. 5.1), the variance is equal to the mean $\mathrm{Var}_{\mathcal{P}}[W] = w$. Because we assumed that each individual reproduces according to an independent distribution, the covariance in fitness between any pair of individuals is zero.

We now investigate the fundamental consequences of variance in individual fitness for evolutionary dynamics.

5.2 Identity-by-descent and loss of diversity

In order to start apprehending the phenomenon of genetic-drift we assume a population of constant size n and that is monomorphic. Constant size may be enforced when the population has reached its carrying capacity under competition for space and assuming that fecundity is large (whereby there will no variation in n), In the presence of a variance in fitness, a population is monomorphic if all individuals have the same expected fitness, the same variance in fitness, the same correlations in fitness, and so on for all higher moments. These assumptions imply that fitness $w = 1$ for all individuals and that we can denote by σ^2 the common variance in fitness (i.e., $\sigma^2 = \mathrm{Var}_{\mathcal{P}}[W_i]$ for any individual $i = 1, 2, ..., n$ from the population). We denote by ρ the covariance in fitness between any pairs of individuals in the population. Importantly, under these assumptions, there is no natural selection in this population, since there is no heritable variance in fitness.

In any finite population of constant size n the variance in fitness of an individual, σ^2, is related to the covariance in fitness, ρ, by the identity

$$\frac{\sigma^2}{n-1} = -\rho. \tag{5.3}$$

[1] Formally, the conditional average $\mathrm{E}_{\mathcal{P}}[W]$ is usually written $\mathrm{E}[W \mid \mathcal{P}]$ (see Appendix 1), but the shorthand notation $\mathrm{E}_{\mathcal{P}}[W]$ is convenient.

This can be understood by noting that if an individual leaves more than one descendant due to chance effects (since on average fitness is one, $w = 1$), then at least one other individual must leave less than one descendant in order to maintain a constant population size. Thus, the variance in fitness of a focal individual, per randomly sampled other individual in the population (there are $n - 1$ other individuals), is equal to the covariance in fitness of the focal individual with that other individual. The take home message behind eq. (5.3) being that in the presence of reproductive variance, some individuals will reproduce more than others just by chance, and that this effect can be quantified.

5.2.1 The coalescence probability

The term $\sigma^2/(n-1)$ in eq. (5.3) is a measure of the extent to which a randomly sampled individual in the population will leave at any point in time more than one descendant, even though on average every individual has the same fitness $w = 1$. Since the right hand side of eq. (5.3) is the negative of the covariance in offspring number between two individuals in the parental generation, the left hand side is equal to the probability Φ that two individuals, randomly sampled in the population in the offspring generation descend from the same individual in the parental generation. That is, we have

$$\Phi = \frac{\sigma^2}{n-1}, \tag{5.4}$$

which is called the **coalescence probability**. Hence, when the variance σ^2 is greater than zero, two individuals may descend from the same ancestor and because the variance σ^2 is generally of lower order of magnitude than total population size, the coalescence probability is of the order of the inverse of the population size. In fact, if we assume that offspring number follows the Poisson distribution (eq. 5.1), then the variance in fitness is $\sigma^2 = 1$ and in a population that is not too small, we can equate the denominator in eq. (5.4) to n as the "1" has virtually no effect, whereby in a large population

$$\Phi = \frac{1}{n}. \tag{5.5}$$

This says that the chance that two offspring come from the same parent is $1/n$, which is intuitive since we have n possible parent that can contribute to the offspring generation.

5.2.2 Stochastic loss of diversity

The coalescence probability Φ allows us to evaluate the probability q_t that in an asexual population two individuals, randomly sampled in the population, are **identical-by-descent**. In other words, q_t is the probability that the two individuals sampled at demographic time period t have a **common ancestor** in the past, i.e., they belong to the same "dynasty" or lineage of individuals initiated by a single individual at time $t = 0$ (see Fig. 5.1). This

satisfies the recurrence equation

$$q_t = \Phi + (1 - \Phi)q_{t-1}, \tag{5.6}$$

since with probability Φ two randomly sampled individuals have the same parent in the previous generation, while with probability $1 - \Phi$ they have two different parents, but these two parental individuals themselves descend from the same ancestor with probability q_{t-1}.

Eq. (5.6) shows that as time passes by (and assuming no mutation) individuals tend to descend from the same ancestor, i.e., they tend to become identical-by-descent. The solution to eq. (5.6) by setting $q_0 = 0$ (all individuals are different in the first generation) is

$$q_t = 1 - (1 - \Phi)^t. \tag{5.7}$$

Since Φ is smaller than one, the only equilibrium $q^* = \lim_{t\to\infty} q_t$ satisfying eq. (5.6) is

$$q^* = 1. \tag{5.8}$$

This means that asymptotically all individuals will descend from the same ancestor, whereby the total population tends to uniformity over time in the absence of mutation (see Fig. 5.1).

An important genetic feature implied by this process of accumulation of identity-by-descent is that, if we are considering a gene where different individuals have initially different alleles, then diversity is lost as time passes by. This means that one allele **drifts** to fixation in the population while all others go extinct, since at the end of the process all individuals (and thus all homologous genes) descend from the same ancestor. We now look explicitly at the change in allele frequency in a population that is subject to random genetic drift.

5.3 Allele frequency change with deterministic and stochastic effects

We now assume, as usual, that only two alleles can segregate in the population, A and B. In order to model the impact of the reproductive variance on allele frequency change, we must treat the frequency p of allele A as a random variable, and denote by $\mathrm{E}_{\mathcal{P}}[p']$ its expected frequency in the offspring generation given parental population state $\mathcal{P}$. With this, we can write the change Δp in frequency of allele A as

$$\Delta p = \underbrace{\Delta p_{\mathrm{deter}}}_{\text{“deterministic effect”}} + \underbrace{\Delta p_{\mathrm{drift}}}_{\text{“stochastic effect”}} \tag{5.9}$$

where

$$\Delta p_{\mathrm{deter}} = \mathrm{E}_{\mathcal{P}}[p'] - p$$

is the difference between the expected frequency in the offspring generation and the parental frequency, and

$$\Delta p_{\mathrm{drift}} = p' - \mathrm{E}_{\mathcal{P}}[p'] \tag{5.10}$$

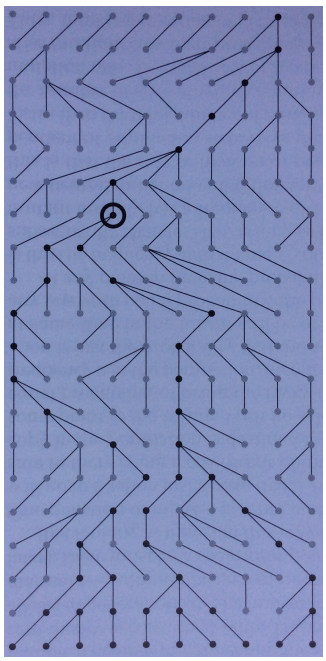

Figure 5.1: Population of constant size exhibiting fixation of a dynasty of individuals owing to demographic stochasticity as time passes by. The most recent common ancestor to all members of the population taken at the bottom of the figure is the individual that is circled. In this figure the probability of coalescence per generation is $\Phi = 1/n$.

is the deviation between the realized frequency p' in the offspring generation and the expected frequency. Eq. (5.9) thus decomposes the change in frequency into a deterministic component, the first term (given by eq. 4.7), and a stochastic component, the second term.

By the decomposition of allele frequency change into deterministic and stochastic terms, the average change du to chance effects is necessarily zero

$$\mathrm{E}_{\mathcal{P}}[\Delta p_{\mathrm{drift}}] = \mathrm{E}_{\mathcal{P}}[p' - \mathrm{E}_{\mathcal{P}}[p']] = \mathrm{E}_{\mathcal{P}}[p'] - \mathrm{E}_{\mathcal{P}}[p'] = 0, \tag{5.11}$$

which means that, on average, allele frequency neither goes up or down. But what about the variance $\mathrm{Var}_{\mathcal{P}}[\Delta p]$ in allele frequency change, which captures allele frequency fluctuations due to stochasticity? Taking the variance over eq. 5.9 and noting that the variance of a constant is zero (see Appendix 1 for a definition of the variance), the variance in allele frequency in the offspring generation is equal to the variance of the stochastic fluctuations

$$\mathrm{Var}_{\mathcal{P}}[\Delta p] = \mathrm{Var}_{\mathcal{P}}[\Delta p_{\mathrm{drift}}] = \mathrm{Var}_{\mathcal{P}}[p']\,. \tag{5.12}$$

This shows that the typical allele frequency fluctuation in eq. (5.9) is the square root of this

variance:

$$\Delta p_{\text{drift}} \underbrace{\pm}_{\text{“on average”}} \sqrt{\text{Var}_{\mathcal{P}}[p']}, \tag{5.13}$$

where this fluctuation can be either positive of negative. Hence, from eq. (5.9) and eq. (5.13), the typical frequency of allele A in the offspring generation is

$$\Delta p = \Delta p_{\text{deter}} \underbrace{\pm}_{\text{“on average”}} \underbrace{\sqrt{\text{Var}_{\mathcal{P}}[p']}}_{\text{“stochastic effect”}} . \tag{5.14}$$

A useful way to quantify the relative importance of stochastic versus deterministic effects on change in allele frequency is the **coefficient of variation**

$$c_{\text{v}} = \frac{\sqrt{\text{Var}_{\mathcal{P}}[p']}}{\text{E}_{\mathcal{P}}[p']}, \tag{5.15}$$

which is the ratio of the standard deviation (the typical size of a stochastic fluctuation) to the mean allele frequency. If the coefficient of variation is close to zero, then stochasticity is negligible and allele frequency change will essentially be equivalent to that under the deterministic model (provided that all higher order deviations, e.g., kurtosis, skewness, are smaller than the variance, which is often the case in practice).

5.3.1 Drift as an evolutionary force

To make these ideas concrete, suppose there is no natural selection nor mutation. Then, the allele frequency change due to deterministic (or directional) effects is nil:

$$\Delta p_{\text{deter}} = 0 \tag{5.16}$$

(recall that in the deterministic process, if we set $s = 0$ and $\mu = 0$ in eq. 4.7 then $\Delta p = 0$). In this case, the change in allele frequency from the parental to the offspring generation will be entirely determined by stochasticity effects whereby using eq. (5.16) in eq. (5.14), we have that

$$\Delta p = \underbrace{\pm}_{\text{“on average”}} \underbrace{\sqrt{\text{Var}_{\mathcal{P}}[p']}}_{\text{“stochastic effect”}} . \tag{5.17}$$

Here, the variance is given by

$$\text{Var}_{\mathcal{P}}[p'] = \Phi p(1-p), \tag{5.18}$$

which is the variance in allele frequency among individuals in the parental generation, $p(1-p)$, times the coalescence probability Φ (see Appendix 3 for a proof). Eq. (5.18) can be understood by first noting that allele frequency will fluctuate only if the variance among individuals changes (since expected allele frequency does not change). Second, the variance among individuals changes only if at least two individuals descend from the same parent, otherwise nothing is affecting this variance.

If we now combine the coalescence probability in eq. (5.5) with eq. (5.18) and substitute into eq. (5.17), then the typical fluctuation in allele frequency is

$$\Delta p \underbrace{\pm}_{\text{“on average”}} \sqrt{p(1-p)} \times \frac{1}{\sqrt{n}}. \tag{5.19}$$

The corresponding coefficient of variation (obtained on substituting eq. 5.16 and eq. 5.18 into eq. 5.15) is

$$c_{\text{v}} = \sqrt{\frac{(1-p)}{p}} \times \frac{1}{\sqrt{n}}. \tag{5.20}$$

The take home message from eqs. (5.19)–(5.20) is twofold:

• Fluctuation of allele frequency due to chance effects will be of the order $1/\sqrt{n}$ under demographic processes where the coalescence probability is of the order of $1/n$. For instance in a population of size $n = 100$, this will result in fluctuations in frequencies of typical order $1/\sqrt{100} = 0.1$. Stochastic effects thus become very small in very large populations because $1/\sqrt{n} \to 0$ as $n \to \infty$ (as long as the variance remains bounded, which is reasonable for biological populations). This is an example of the law of large numbers. This says that if the same (stochastic) experiment is repeated a large number of times (here the experiment is the reproduction of an individual, which is repeated since we consider a population of individuals), then the result should come close to the expected value for the experiment (here the expected reproduction), which in terms of allele frequency dynamics is equivalent to the deterministic process.

• The typical magnitude Δp of allele frequency fluctuations depends on the allele frequency. The lower the allele frequency, the less number of individuals carrying the allele and thus the stronger the fluctuations, which is why the fluctuation goes up as the frequency decreases ($\sqrt{\frac{(1-p)}{p}}$ increases as p decreases). The maximal fluctuations is obtained when there is only one allele in the population, i.e., $p = 1/n$, in which case the coefficient of variation becomes $c_{\text{v}} = \sqrt{\frac{n-1}{n}}$, which is essentially equal to one for large populations [i.e., $(n-1)/n \to 1$ as n becomes large).

The fundamental evolutionary consequence of eqs. (5.19)–(5.20) is that, even if there is no directional evolutionary force on A (no mutation nor selection), there will be fluctuations in p due to reproductive variance and on average the magnitude of these fluctuations will be of the order $1/\sqrt{n}$ (more generally of order $1/\sqrt{\Phi}$). As these fluctuations are as likely to be positive or negative, the frequency p will drift around and allele A will eventually either go to fixation or become extinct in the population. This is so because for any non-zero frequency $p \neq 0$, there must be some change in frequency upwards or downwards (eq. 5.18). This is the process of random genetic drift displayed in Fig. (5.2), which leads to the *survival of the luckiest* or the so-called *neutral evolution*. Because this process applies equally to any allele in the population, the population will ultimately become monomorphic regardless of its initial state.

5.3.2 The interaction between selection and drift

Suppose now that natural selection occurs and that it does so through constant selection such that $w_{\mathrm{B}} = 1$ and $w_{\mathrm{A}} = 1 + s$ in eq. (3.9). Then, the change in the frequency of A in the absence of mutation is that of the deterministic process

$$\Delta p_{\mathrm{deter}} = \frac{p(1-p)s}{1+sp}, \tag{5.21}$$

where s is the (constant) selection advantage of A. In order to get a sense of the biological conditions under which stochasticity interacts with selection, we need to evaluate the coefficient of variation and thus need an expressions for $\mathrm{E}_{\mathcal{P}}[p']$ and $\mathrm{Var}_{\mathcal{P}}[p']$. Since, $\Delta p_{\mathrm{deter}} = \mathrm{E}_{\mathcal{P}}[p'] - p$, we have from eq. (5.21) that the expected frequency in the offspring generation is

$$\mathrm{E}_{\mathcal{P}}[p'] = \frac{(1+s)\,p}{1+sp} \tag{5.22}$$

(recall also eq. 3.11). Now, $\mathrm{Var}_{\mathcal{P}}[p']$ will in general be complicated to evaluate exactly. But if both natural selection and genetic drift are of small magnitude (small s and small Φ), we can directly use $\mathrm{Var}_{\mathcal{P}}[p'] = \Phi p(1-p)$ (eq. 5.18). Substituting this along with eq. (5.22) into eq. (5.15), which leads to the coefficient of variation

$$c_{\mathrm{v}} = \sqrt{\frac{1-p}{p}} \times \frac{1}{\sqrt{n}} \times \left(\frac{1+ps}{1+s}\right), \tag{5.23}$$

where we used the widespread value $\Phi = 1/n$ for the coalescence probability (eq. 5.5). This expression for the coefficient of variation is more complicated, but it says two fundamental things about the interaction between selection and drift:

• When allele A has an appreciable frequency (say $p > 0.1$), the term $\sqrt{(1-p)/p}$ will not be large (< 3). Because of the second term in eq. (5.23), the coefficient of variation is proportional to $1/\sqrt{n}$ and will thus become small in large populations, such that natural selection will generally dominate genetic drift. As a rule of thumb, the outcome of evolution will be mainly determined by natural selection if the coefficient of selection s is much larger than the inverse of population size; namely, if

$$s \gg 1/n. \tag{5.24}$$

• When allele A is rare, say $p = 1/n$ because it has arisen as a single copy by mutation, then the term $\sqrt{(1-p)/p} \times (1/\sqrt{n}) = \sqrt{\frac{n-1}{n}} \approx 1$ is much larger (essentially equal to one as we saw in section 5.3.1). At this point, natural selection becomes a very weak evolutionary force that is likely to be of lesser importance compared to genetic drift, owing to the fact that stochasticity alone is the main driver of allele frequency change. This implies that even when an allele is favored by selection, it can go extinct by chance effects when it is rare and this holds regardless of population size. Conversely, an allele that is counter-selected ($s < 0$) may actually increase in frequency due to chance effects; in this way deleterious alleles can increase in frequency.

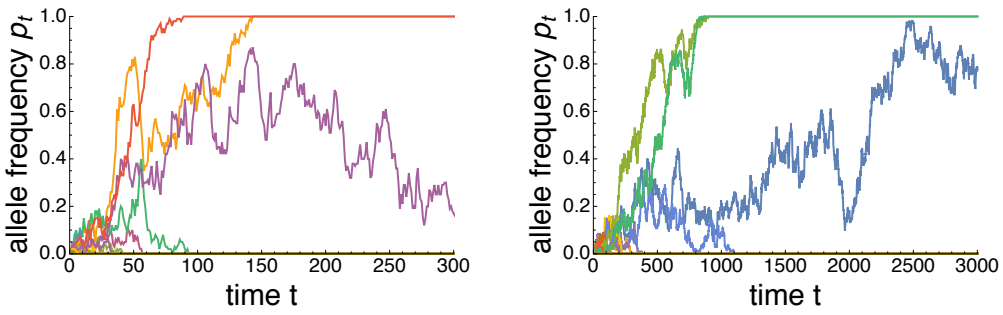

Figure 5.2: Frequency p_t of allele A as a function of time t in a population of constant size n that is only subject to random genetic drift (where the coalescence probability is $1/n$). The initial number of mutants is one ($p_0 = 1/n$). In the left panel populaiton size is $n = 100$ and 100 runs are displayed (each line is one run), while in the right panel population size is $n = 1000$ and 1000 runs are displayed. As can be seen from the two panels, for most of the runs allele A goes extinct, while it increases in frequency and reaches fixation by drift in a few number of cases.

The interaction between natural selection and genetic drift thus depends on the frequency p of allele A in the population. For any novel allele appearing as a single copy, drift becomes essentially as strong as natural selection. Hence, a mutant allele must cross a threshold frequency to be favored by natural selection. Once this threshold has been crossed, stochastic effects become negligible and natural selection becomes the main evolutionary force.

5.4 Appendix

Appendix 1. Here, we briefly recall some notions of probability theory. In particular, the expectation of a discrete random variable X is defined as

$$\mathrm{E}[X] = \sum_x x \Pr(X = x), \tag{5.25}$$

where $\Pr(X = x)$ is the probability that the random variable takes the realized value x. The expectation is thus the average (or mean) of the random variable X. The variance of a random variable is

$$\mathrm{Var}[X] = \mathrm{E}\big[(X - \mathrm{E}[X])^2\big], \tag{5.26}$$

which is the expectation of the squared deviation relative to the mean. Since the conditional probability $\Pr(X = x \mid Y)$ that event x occurs given event Y is a probability,

we can define the conditional expectation of the random variable X given event Y as

$$\mathrm{E}[X \mid Y] = \sum_x x \Pr(X = x \mid Y), \tag{5.27}$$

and the same logic follows for the conditional variance $\mathrm{Var}[X \mid Y]$.

Appendix 2. We here derive eq. (5.4). Let us denote by W_i the random number of offspring by individual i from a parental population (hence we index individuals by $i = 1, 2, 3, ...n$). The coalescence probability Φ that two individuals, randomly sampled in the offspring generation, descend from the same individual in the parental generation is given by

$$\Phi = \mathrm{E}_{\mathcal{P}}\left[\sum_{i=1}^{n} \frac{W_i(W_i-1)}{n(n-1)}\right] = \sum_{i=1}^{n} \frac{\mathrm{E}_{\mathcal{P}}[W_i(W_i-1)]}{n(n-1)} = \frac{\sigma^2}{n-1}. \tag{5.28}$$

This is the expectation of the ratio of the total number of ways of sampling two individual from the same parent to the total number of ways of sampling two individuals, given the parental population state $\mathcal{P}$, which is of size n. The second equality follows from the linearity of expectations and the third equality follows from the fact that $\mathrm{E}_{\mathcal{P}}[W_i(W_i-1)] = \sigma^2 + \mathrm{E}_{\mathcal{P}}[W_i]^2 - \mathrm{E}_{\mathcal{P}}[W_i] = \sigma^2$, since $\mathrm{E}[W_i] = w = 1$ as all individuals are the same (they have the same variance, same mean, etc.) and fitness is on average 1.

Appendix 3. Here we prove eq. (5.18). For this, we first recall that owing to the property of the variance we can write

$$\mathrm{Var}_{\mathcal{P}}\left[p'\right] = \mathrm{E}_{\mathcal{P}}\left[(p' - \mathrm{E}_{\mathcal{P}}\left[p'\right])^2\right] = \mathrm{E}_{\mathcal{P}}\left[(p')^2\right] - \mathrm{E}_{\mathcal{P}}\left[p'\right]^2. \tag{5.29}$$

Further, we have

$$\mathrm{E}_{\mathcal{P}}\left[(p')^2\right] = \Phi\,\mathrm{E}_{\mathcal{P}}[p] + (1-\Phi)\,\mathrm{E}_{\mathcal{P}}\left[p^2\right], \tag{5.30}$$

since $\mathrm{E}_{\mathcal{P}}\left[(p')^2\right]$ is the probability of sampling in the offspring generation two individuals that carry allele A. Two individuals in the offspring generation carry A if they descend from the same parent (probability Φ) and that parent carries A (probability $\mathrm{E}_{\mathcal{P}}[p]$) or if the two individuals descend from two distinct parents (probability $1-\Phi$) that both carry A (probability $\mathrm{E}_{\mathcal{P}}\left[p^2\right]$). Since the expectation is conditional on the parental generation that has frequency p of allele A, we have $\mathrm{E}_{\mathcal{P}}[p'] = \mathrm{E}_{\mathcal{P}}[p] = p$ and $\mathrm{E}_{\mathcal{P}}\left[p^2\right] = p^2$, whereby substituting this along with eq. (5.30) into eq. (5.29), we obtain by simplifying terms that

$$\mathrm{Var}_{\mathcal{P}}\left[p'\right] = \Phi p(1-p). \tag{5.31}$$
