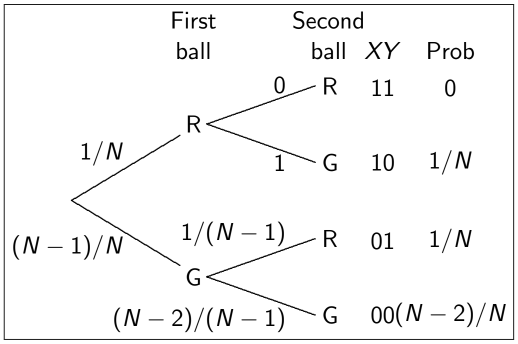
Homework problem 1.41

Imagine an urn with N balls, of which 1 is red and the rest are black. You draw 2 balls from the urn at random *without* replacement. Let $X = 1$ if the first ball is red and $X = 0$ otherwise. Define Y similarly for the second ball.

What is the covariance of X and Y?

Probability tree & distribution of (X, Y)



Goal: calculate $C(X, Y) = E[XY] - E[X]E[Y]$

$E[X]$

X	Y	Pr
1	1	0
1	0	$1/N$
0	1	$1/N$
0	0	$(N-2)/N$

$$\begin{aligned} E[X] &= 1 \times 0 \\ &\quad + 1 \times 1/N \\ &\quad + 0 \times 1/N \\ &\quad + 0 \times (N-2)/N \\ &= 1/N \end{aligned}$$

$E[Y]$

X	Y	Pr
1	1	0
1	0	$1/N$
0	1	$1/N$
0	0	$(N-2)/N$

$$\begin{aligned} E[Y] &= 1 \times 0 \\ &\quad + 0 \times 1/N \\ &\quad + 1 \times 1/N \\ &\quad + 0 \times (N-2)/N \\ &= 1/N \end{aligned}$$

$E[XY]$

X	Y	XY	Pr
1	1	1	0
1	0	0	$1/N$
0	1	0	$1/N$
0	0	0	$(N-2)/N$

$$\begin{aligned} E[XY] &= 1 \times 0 \\ &\quad + 0 \times (\text{the rest}) \\ &= 0 \end{aligned}$$

Covariance of X and Y:

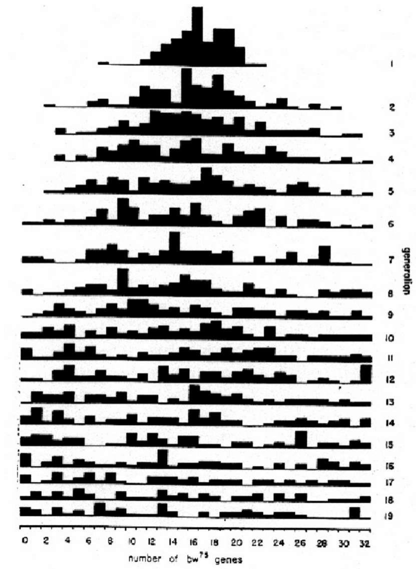
$$\overbrace{E[XY]}^{0} - \overbrace{E[X]}^{1/N}\overbrace{E[Y]}^{1/N} = -1/N^2$$

How Drift Affects Heterozygosity

Alan R. Rogers

January 4, 2024

We begin with data from an experiment, described by Peter Buri in 1956. (Gene frequency in small populations of mutant *Drosophila*, Evolution, 10:367–402)

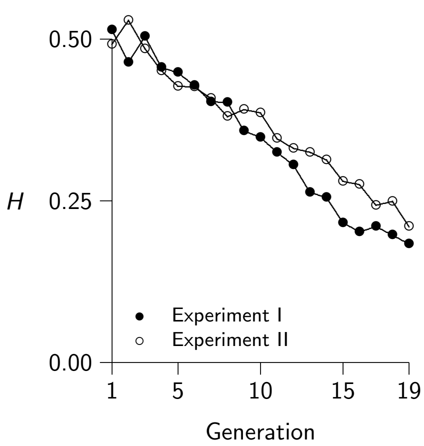


Buri's drift experiment I

- Each generation: 107 bottles, each w/ 8 male & 8 female fruit flies.
- Generation 0: all flies heterozygous.
- Rows show distribution of allele frequency in 19 successive generations.

Peter Buri, 1956

Decay of heterozygosity in Buri's two experiments



- Heterozygosity (H) starts at 0.5
- Declines to about 0.2
- Why?

As heterozygosity declined w/i bottles, the variance among them increased

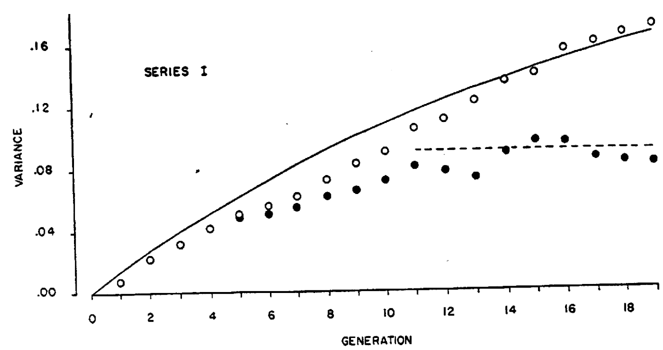
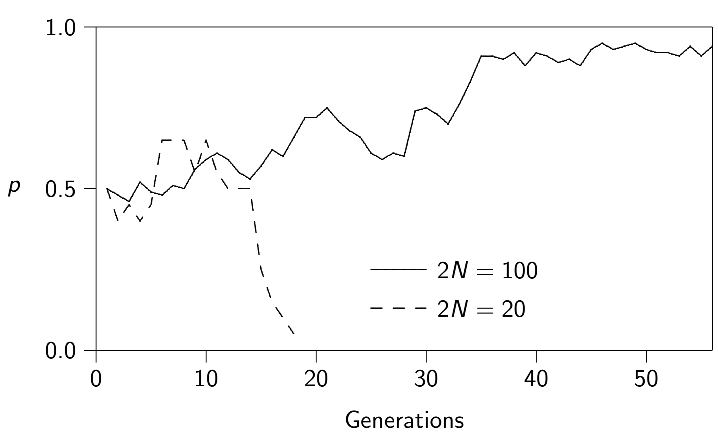


Fig. 12. Theoretical variances of the total frequency distribution by generation including fixed classes and based on a common estimate of $2N_e = 18$ for series I are represented by the smooth curves. Open circles show the observed variance of the distribution including previously fixed classes. Closed circles indicate the observed total variance excluding fixed classes. The asymptote ($= 0.091$) indicates approximately the theoretical maximum value of this variance. All values are on a relative scale.

Computer Simulations of Genetic Drift



The Urn Metaphor

Imagine two urns: metaphors for a population in two successive generations. Urn 1 has 50 balls, some red, some white, representing parental gene copies. Urn 2 is empty until urn 1 has "reproduced" as follows:

1. Examine a random ball from urn 1.
2. Put a ball of the same color into urn 2.
3. Replace the ball from urn 1.
4. Repeat until there are 50 balls in urn 2.

Urn 2 is differs from urn 1 because of random sampling: a metaphor for genetic drift.

The urn model behaves a lot like genetic drift in real populations:

1. variation between populations increases
2. variation within populations decreases

Yet real organisms don't reproduce as our urns do. The best urn model is unlikely to be one in which the number of balls matches the number of gene copies.

Decay of Heterozygosity: Notation

N = # of diploid individuals in population

$2N$ = # of gene copies in population

$\mathcal{G}$ = Probability that two random gene copies, drawn with replacement from generation t, are copies of the same allele.

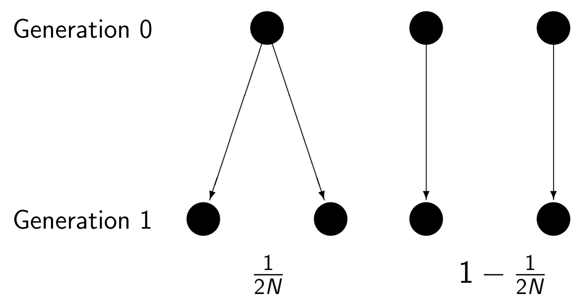
$\mathcal{G}'$ = same thing in the generation $t+1$.

Decay of Heterozygosity: Logic

Two gene copies may be identical in state either because

1. they are copies of the same parental gene copy, or
2. they are copies of distinct parental gene copies, which happen to be identical in state.

Two gene copies either are or are not copies of the same parental gene copy



Two gene copies are copies of the same parental gene copy with probability $1/2N$, and of distinct parental gene copies with probability $1 - 1/2N$.

Event	Prob
Individual carries 2 copies of same parental gene copy	$1/2N$

Explanation:

1. First draw a random gamete from among those produced by the parental generation. This gamete is equally likely to have been produced by any of the $2N$ parental gene copies.
2. Next draw another gamete at random. There is 1 chance in $2N$ that the second is a copy of the same parental gene copy as the first.

Event	Prob
Individual carries copies of 2 distinct parental gene copies, which are themselves identical.	$(1 - 1/2N)\mathcal{G}$

Explanation:

1. The two random gene copies are copies of distinct parental genes with probability $1 - 1/2N$.
2. These distinct parental gene copies are copies of the same allele with probability $\mathcal{G}$—that is the definition of $\mathcal{G}$.
3. *Both* things are true with probability:

$$\left(1 - \frac{1}{2N}\right)\mathcal{G}$$

In short, the two genes are identical if they are copies either of

1. the same parental gene copy (probability $1/2N$), or of
2. distinct but identical gene copies (probability $(1 - 1/2N)\mathcal{G}$).

Altogether,

$$\mathcal{G}' = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\mathcal{G}$$

To see where this goes, it is easier to work with the probability that the two gene copies are copies of *different* alleles, i.e. with the heterozygosity,

$$\begin{aligned}\mathcal{H}' &= 1 - \mathcal{G}' \\ &= \left(1 - \frac{1}{2N}\right)\mathcal{H} \qquad \text{(after some algebra).}\end{aligned}$$

Can you supply the algebra?

The Time-path of Heterozygosity

$$\begin{aligned}\mathcal{H}_1 &= \left(1 - \frac{1}{2N}\right)\mathcal{H}_0 \\ \mathcal{H}_2 &= \left(1 - \frac{1}{2N}\right)\mathcal{H}_1 \\ &= \left(1 - \frac{1}{2N}\right)^2\mathcal{H}_0 \\ \mathcal{H}_t &= \left(1 - \frac{1}{2N}\right)^t\mathcal{H}_0\end{aligned}$$

where $\mathcal{H}_0$ is the original heterozygosity and $\mathcal{H}_t$ is the heterozygosity in generation t.

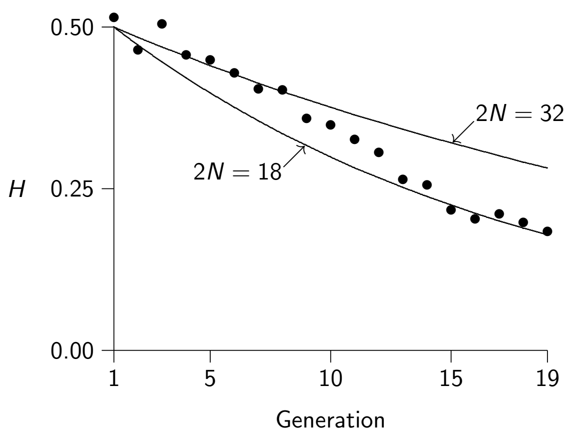
Example

In Peter Buri's experiment, $\mathcal{H}_1 = 1/2$ because half the population were heterozygotes after the first generation of random mating. 18 generations later:

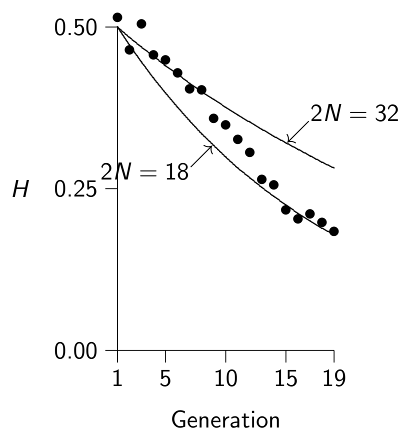
$$\mathcal{H}_{19} = \frac{1}{2}\left(1 - \frac{1}{2N}\right)^{18}$$

But what is $2N$?

Heterozygosity: Buri's experiment I vs. urn model

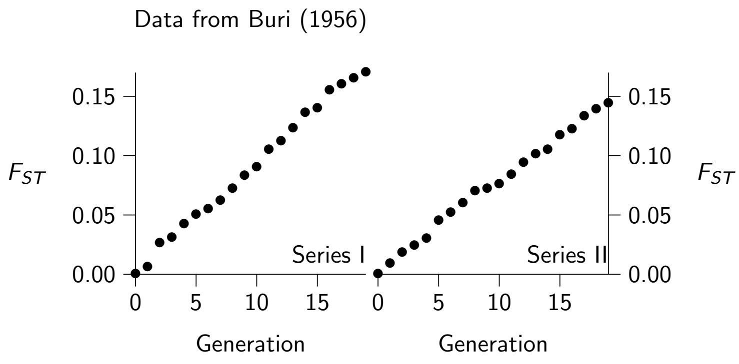


Heterozygosity: Buri's experiment I vs. urn model



- There were 32 gene copies in each bottle.
- Yet $2N = 32$ provides a poor fit to data.
- Better fit with $2N = 18$.
- 18 is the "*effective population size*"

F_{ST} measures variation among populations



Model fits after setting $N = N_e$