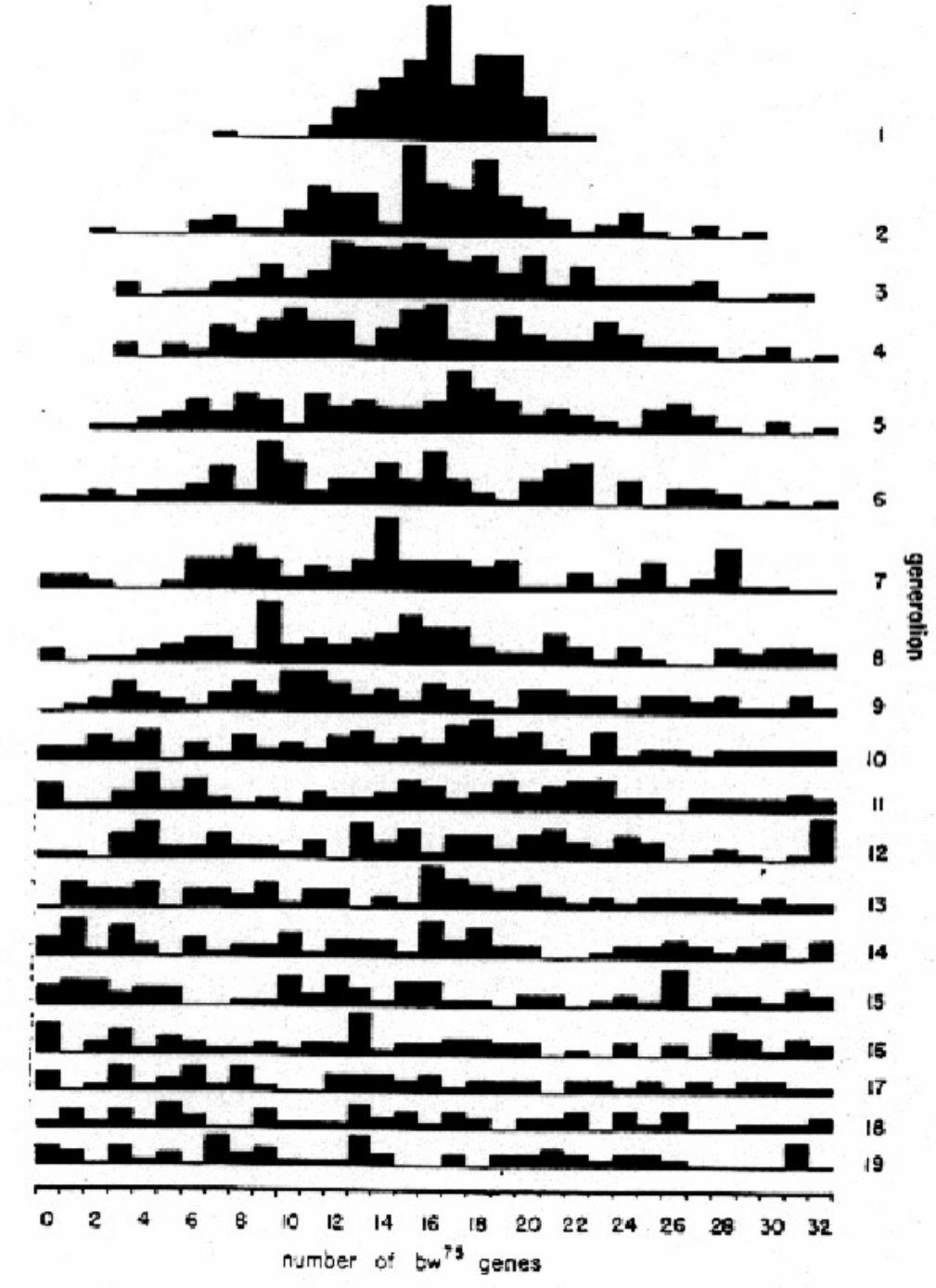
How Drift Affects Heterozygosity

Alan R. Rogers

January 29, 2021

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)

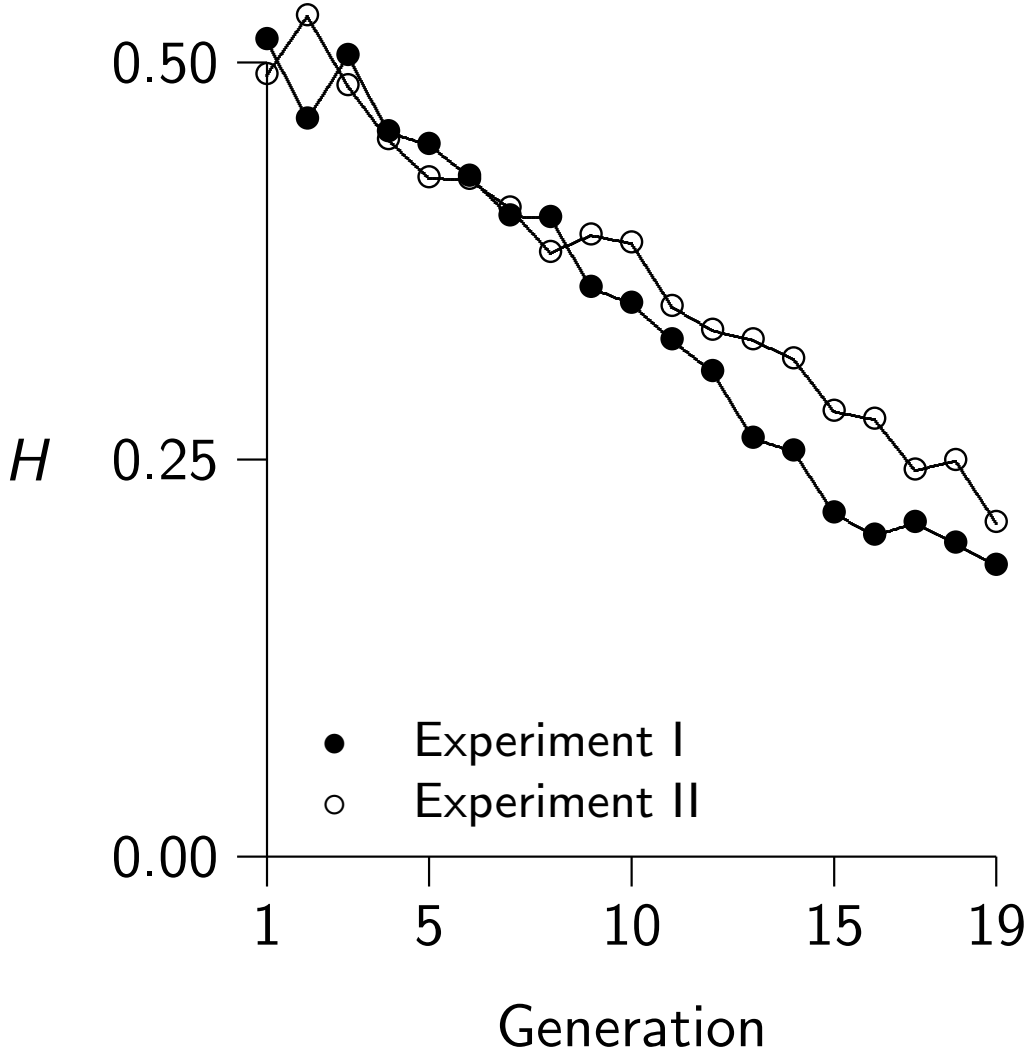


Buri's drift experiment I

- 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

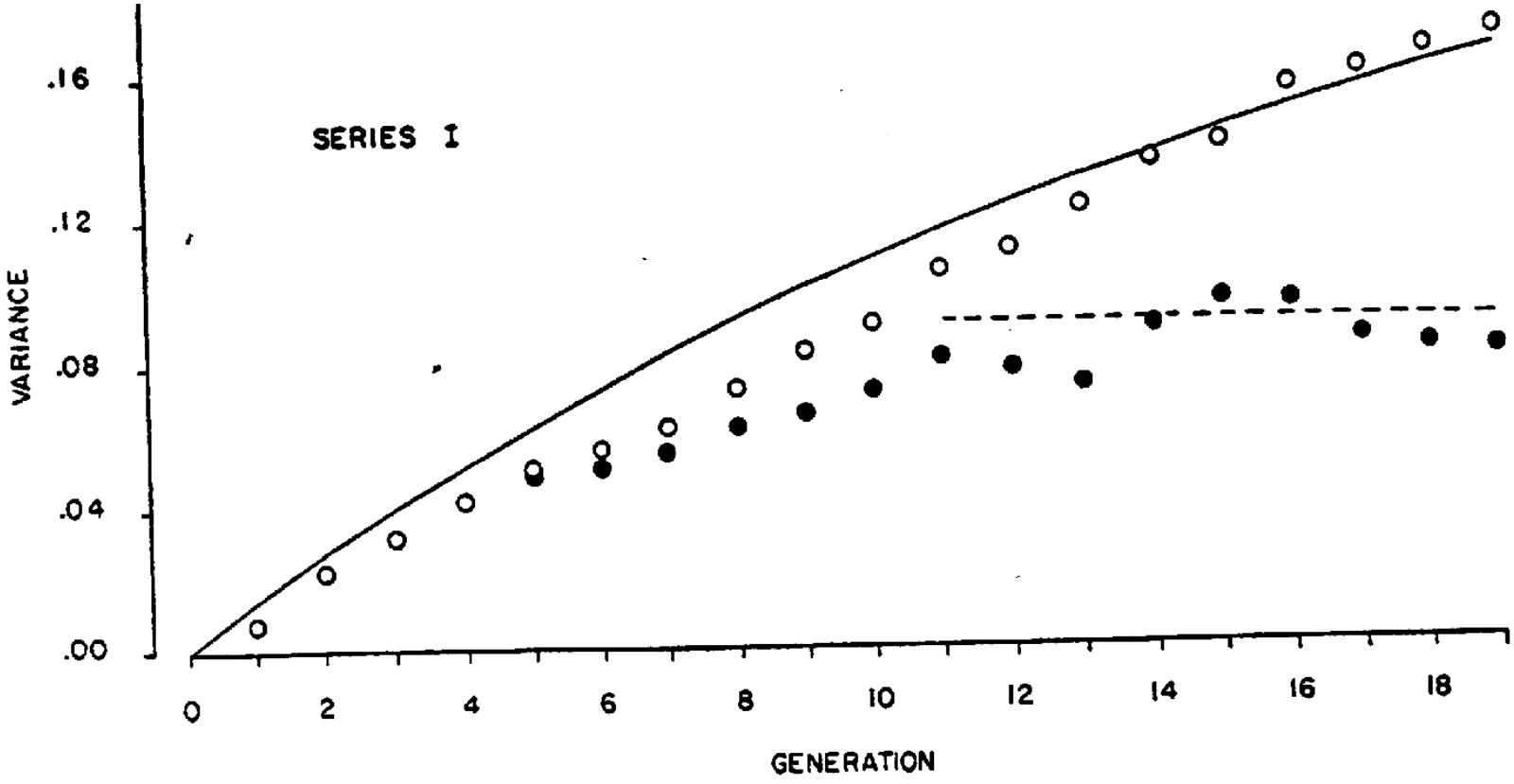
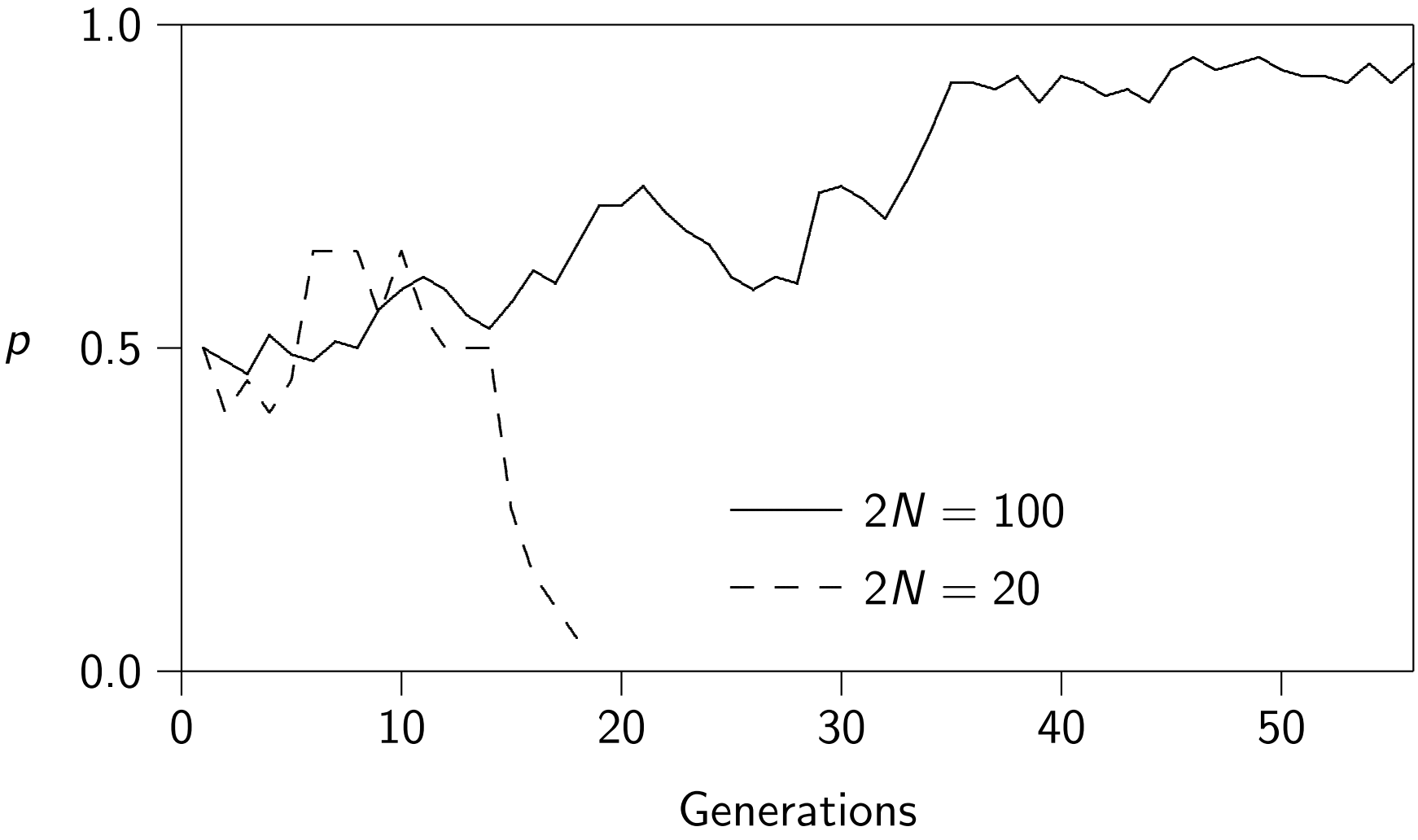


FIG. 12. Theoretical variances of the total frequency distribution by generation including fixed classes and based on a common estimate of $2N_e = 18$ for series I are represented by the smooth curves. Open circles show the observed variance of the distribution including previously fixed classes. Closed 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.

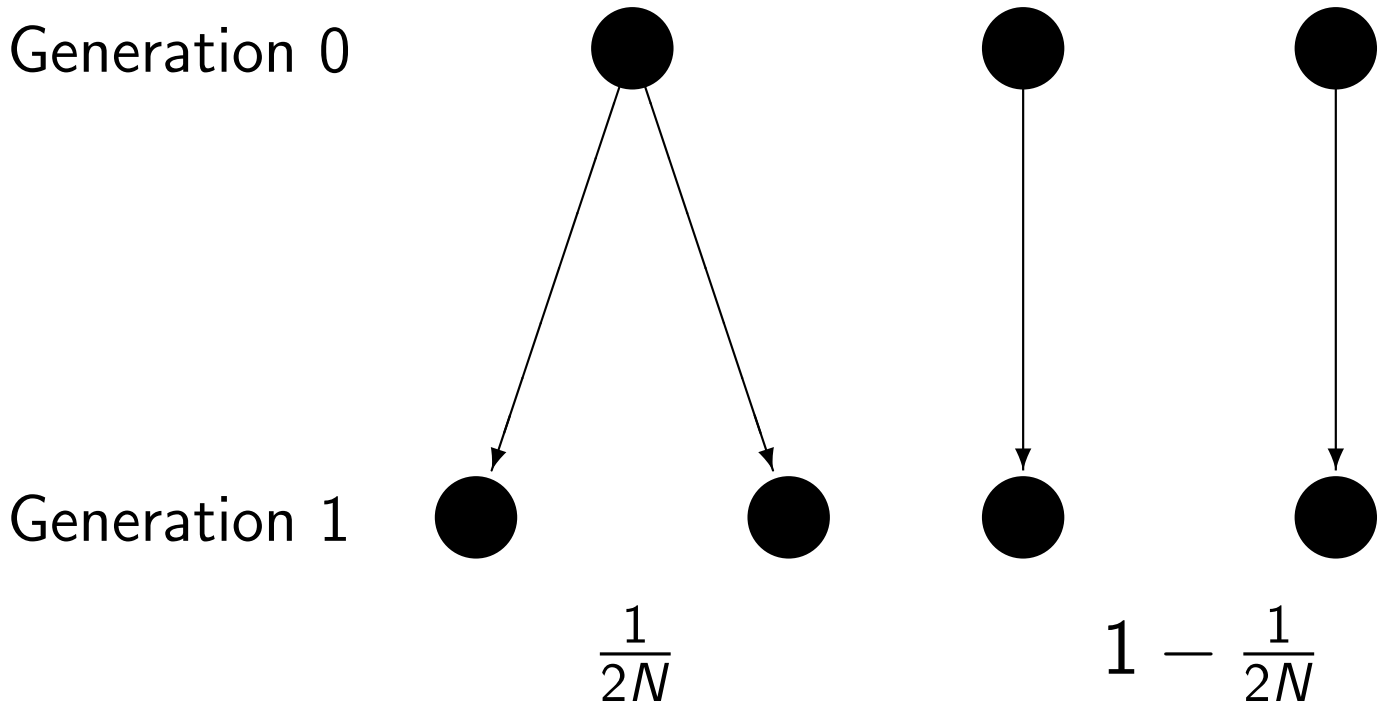
$\mathcal{G}' =$ same thing in the generation $t + 1$.

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

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

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 genes.
2. Next draw another gamete at random. There is 1 chance in $2N$ that the second is a copy of the same parental gene as the first.

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

Explanation:

1. The two random gene are copies of distinct parental genes with probability $1 - 1/2N$.
2. These distinct parental genes 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 (probability $1/2N$), or of
2. distinct but identical genes (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 genes 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.

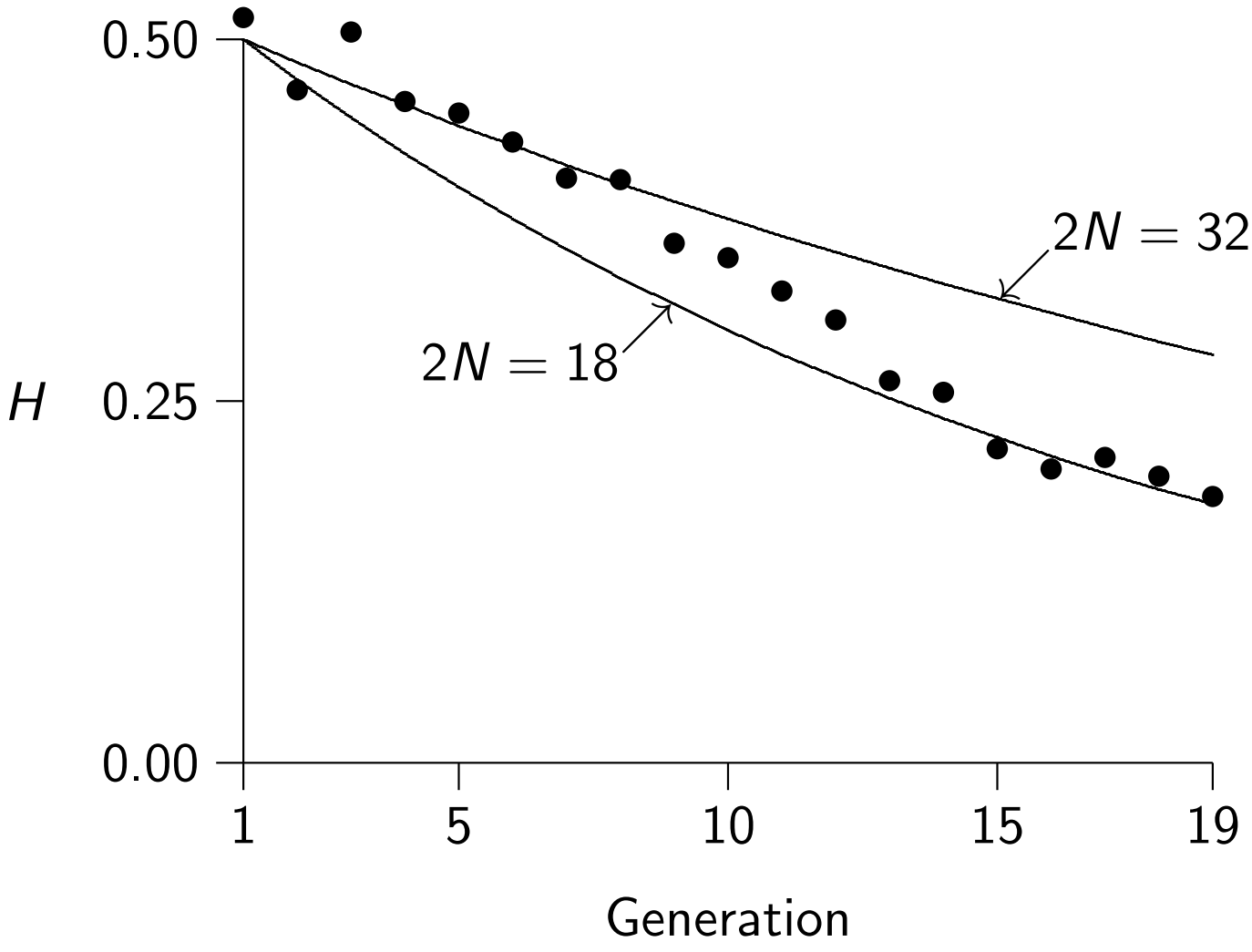
Example

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:

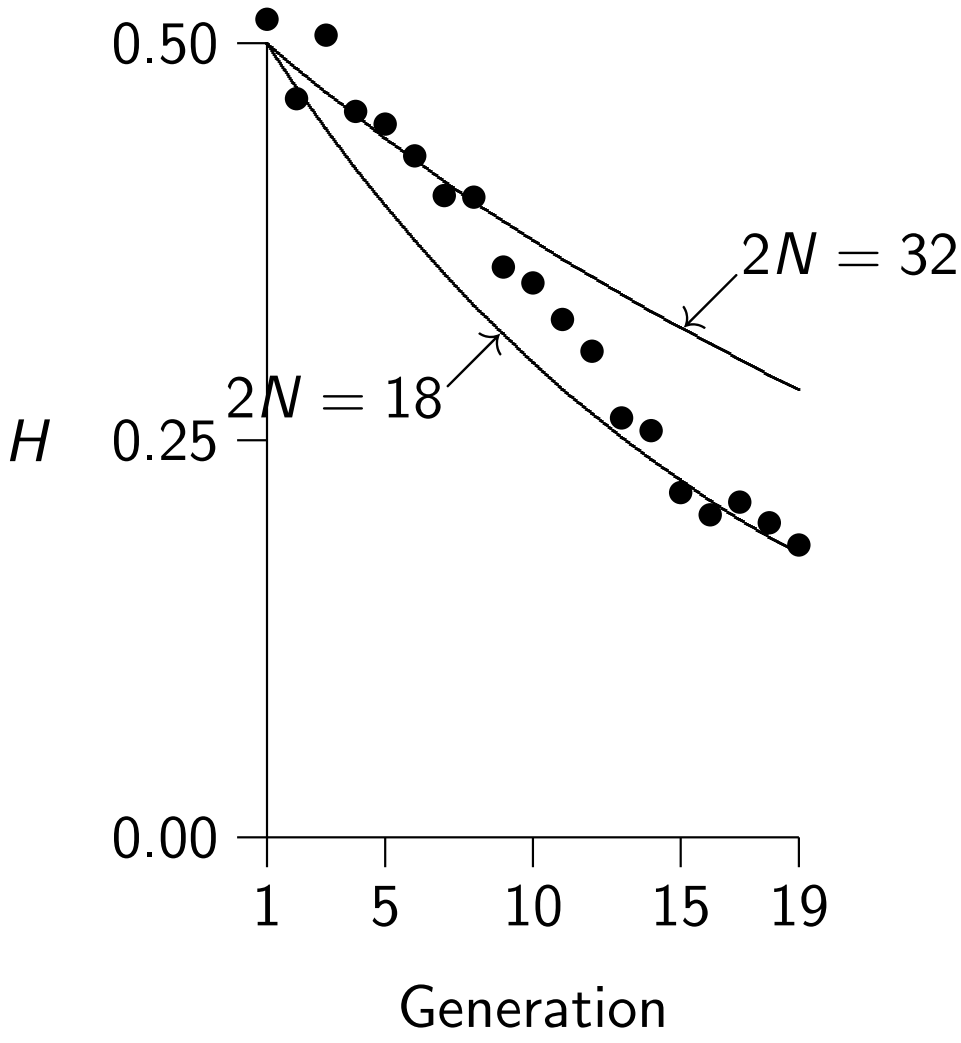
$$\mathcal{H}_{19} = \frac{1}{2}\left(1 - \frac{1}{2N}\right)^{18}$$

But what is $2N$?

Heterozygosity: Buri's experiment I vs. urn model

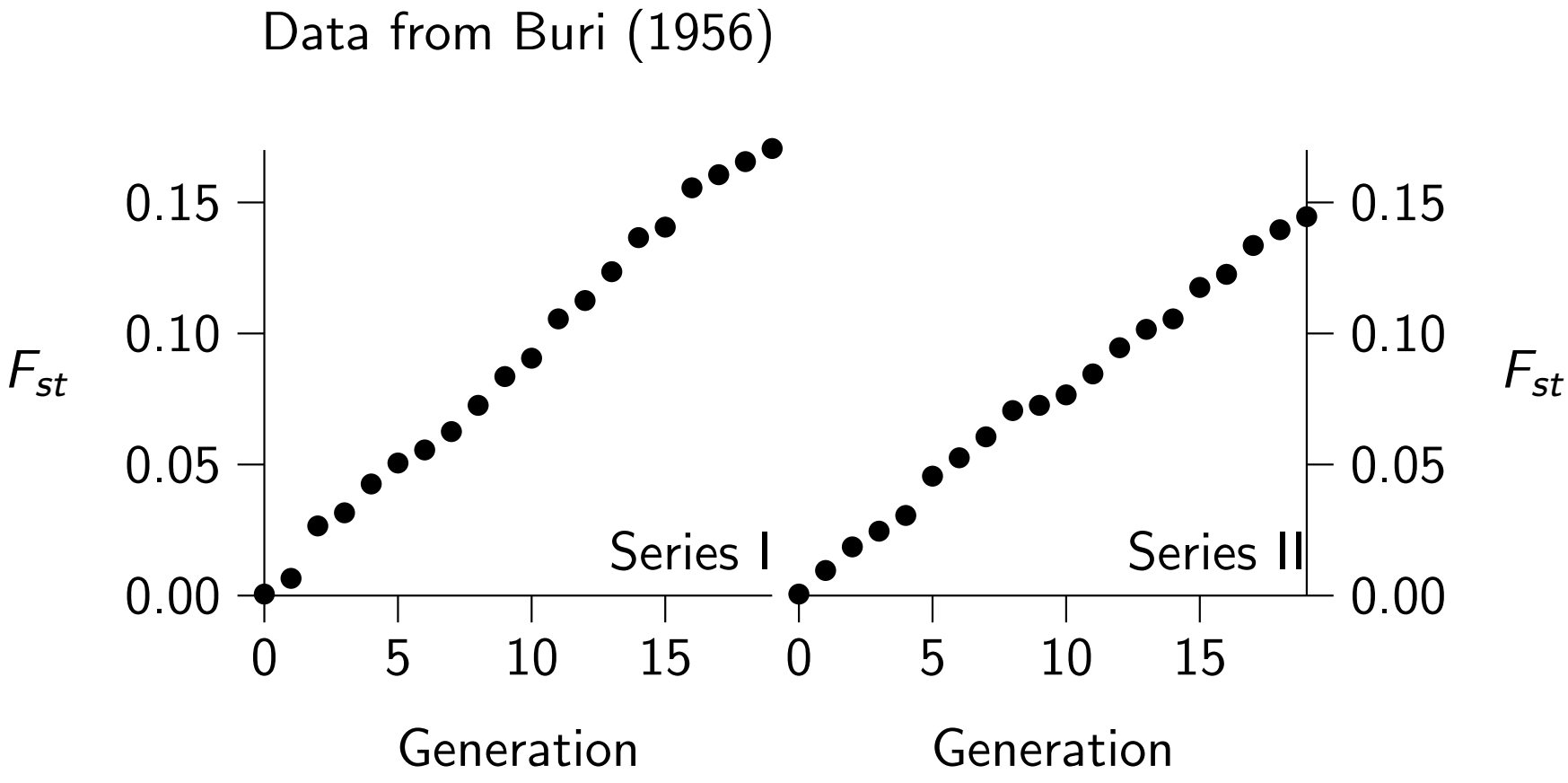


Heterozygosity: Buri's experiment I vs. urn model

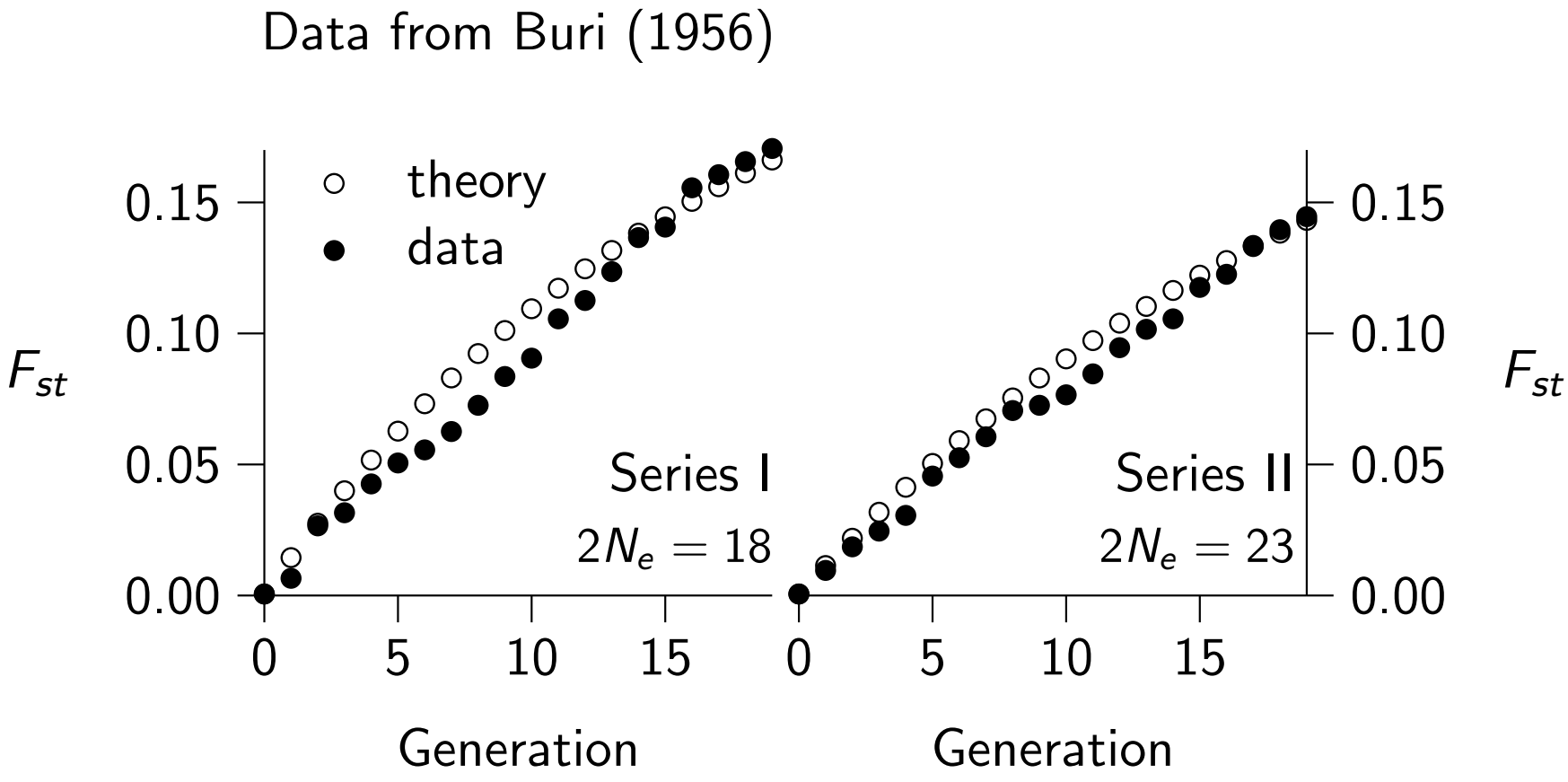


- There were 32 gene copies in each bottle.
- Yet $2N = 32$ provides a poor fit to data.
- Better fit with $2N = 18$.
- 18 is the "*effective population size*"

F_{ST} measures variation among populations



Model fits after setting $N = N_e$



How Mutation Affects the Decay of Heterozygosity

Alan R. Rogers

January 29, 2021

Two models of mutation

The mutation rate is u per gamete per generation.

Infinite alleles	Each mutant is an allele never seen before.
K alleles	When allele i mutates, the mutant is equally likely to be any allele other than i. There are $K-1$ possibilities, each with probability $1/(K-1)$.

We'll focus on the model of infinite alleles.

Model of infinite alleles

- Each mutation creates a unique mutation, which has never been seen before.
- Two identical gene copies remain identical in next generation only if neither mutates.
- Probability of this is $(1 - u)^2$, where u is the mutation rate.

How drift affects gene identity

Without mutation

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

With mutation

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

Approximations

$$\begin{aligned} (1-u)^2 &\approx 1-2u \\ \frac{1-2u}{2N} &\approx \frac{1}{2N} \end{aligned}$$

Numerical example: $(1-u)^2 \approx 1-2u$

u	$(1-u)^2$	$1-2u$
0.100000	0.810000	0.800000
0.010000	0.980100	0.980000
0.001000	0.998001	0.998000
0.000100	0.999800	0.999800
0.000010	0.999980	0.999980
0.000001	0.999998	0.999998

Numerical example: $(1-u)/2N \approx 1/2N$

u	$2N$	$(1-2u)/2N$	$1/2N$
0.0001	10	0.0999800	0.10000
0.0001	100	0.0099980	0.01000
0.0001	1000	0.0009998	0.00100
0.0001	10000	0.0001000	0.00010
0.0001	100000	0.0000100	0.00001

Before approximations

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

After

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

At equilibrium $\mathcal{G}' = \mathcal{G}$, so

$$\begin{aligned} \hat{\mathcal{G}} &= \frac{1}{4Nu+1} \\ \hat{\mathcal{H}} &= 1 - \hat{\mathcal{G}} = \frac{4Nu}{4Nu+1} \end{aligned}$$

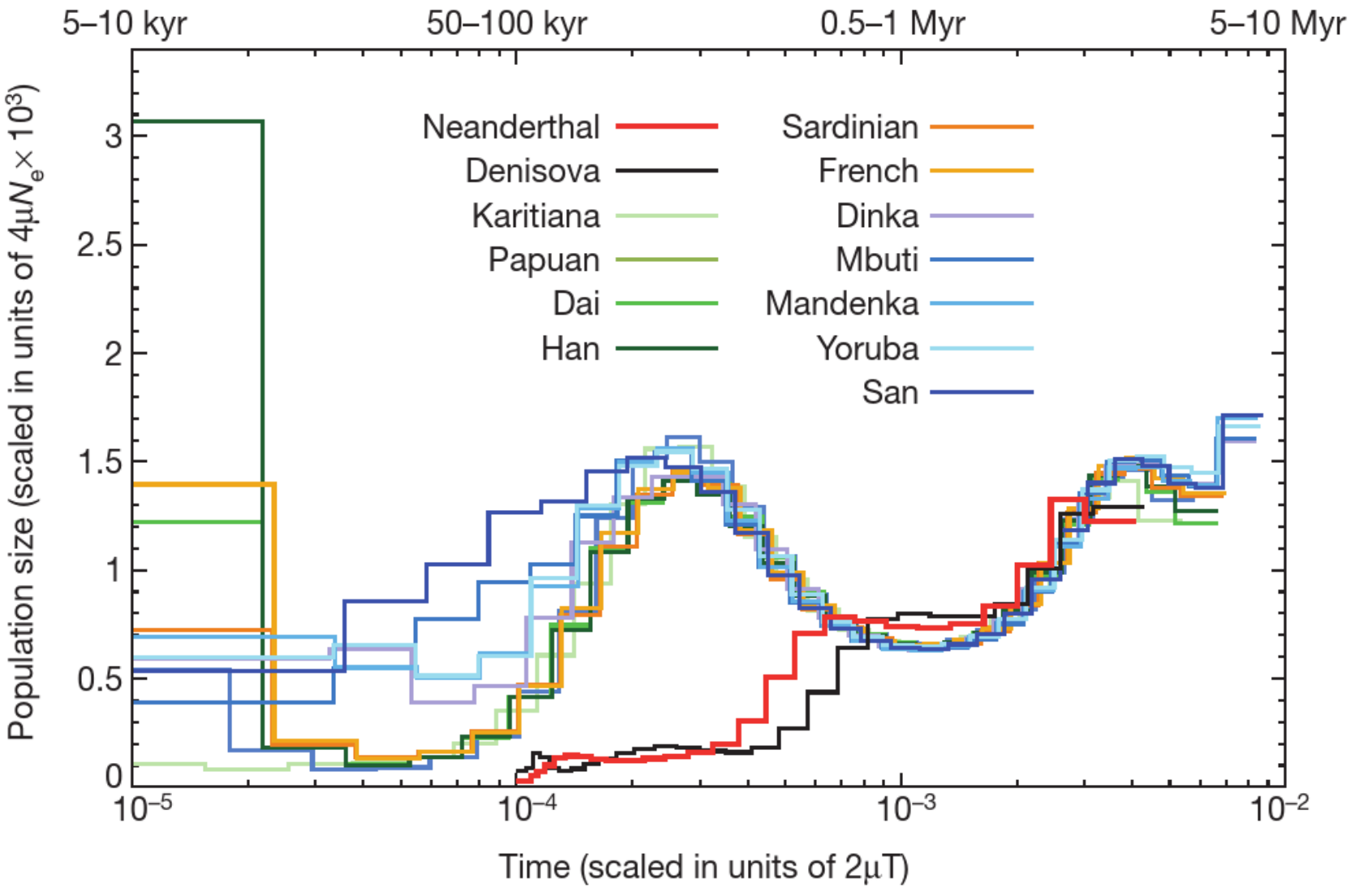
for model of infinite alleles. The "hats" indicate that these are equilibrium values. If $4Nu$ is large, $\hat{\mathcal{H}} \approx 1$.

Model of 2 alleles

$$\hat{\mathcal{H}} = \frac{4Nu}{8Nu + 1}$$

If $4Nu$ is large, $\hat{\mathcal{H}} \approx 1/2$.

History of human population size



Neanderthals had low heterozygosity

Species	Population	Heterozygosity
Neanderthal	El Sidrón	0.000143
	Vindija	0.000127
	Altai	0.000113
Modern	African	0.000507
	European	0.000387
	Asian	0.000358

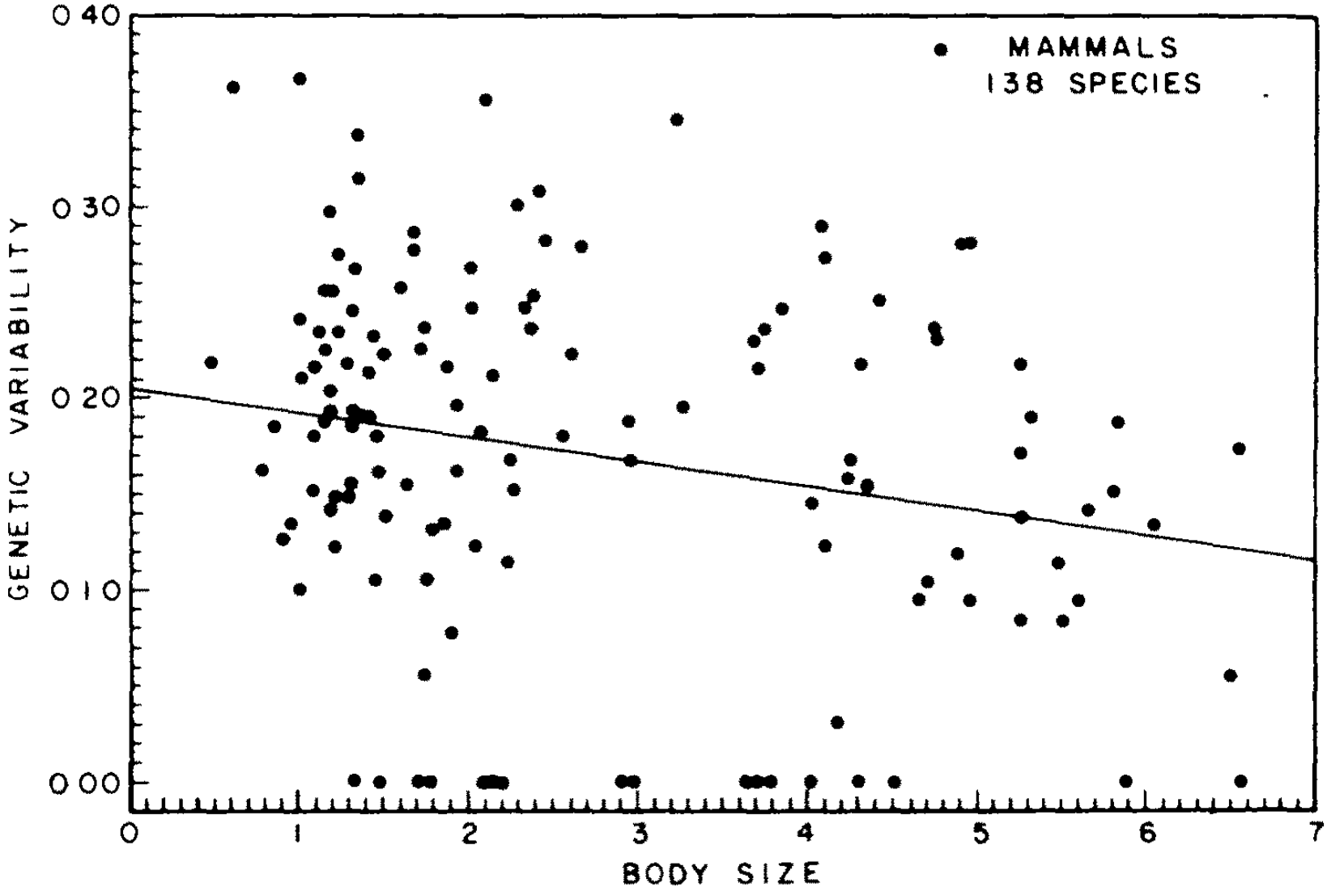
Low heterozygosity $\Rightarrow$ small population.

The magnitude of predicted effects on heterozygosity at a biallelic locus

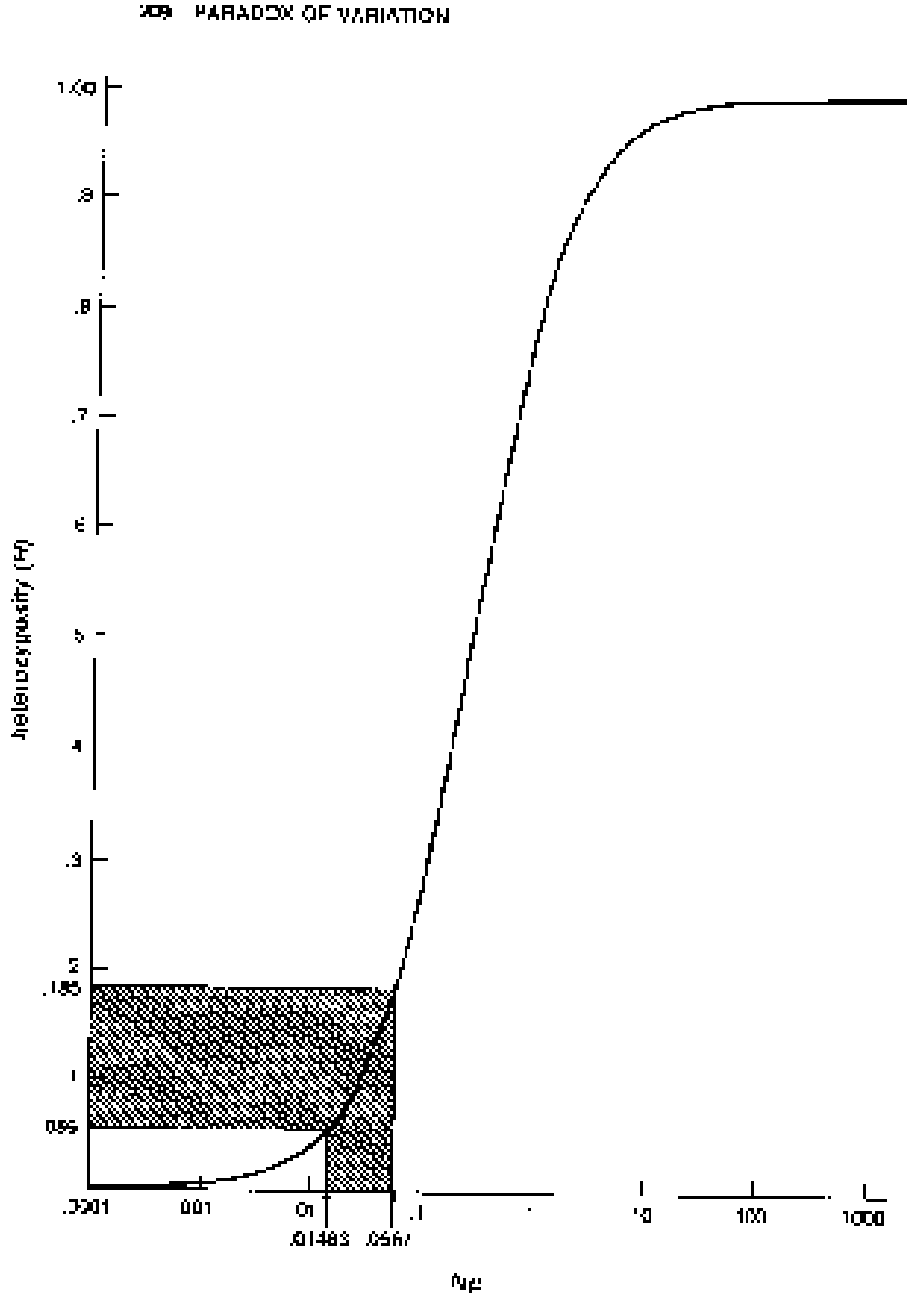
N	H
1,000	0.00004
10,000	0.00040
100,000	0.00400
1,000,000	0.03700

$u = 10^{-8}$

Populations of different size should differ enormously in heterozygosity.
Is this really true?



Heterozygosity at enzyme polymorphisms vs $\log_{10}$ grams of body wgt (Wooten & Smith 1985). Range: 3 g to 4000 kg.



Variation in H implies implausibly-small variation in $N\mu$.

Lewontin (1970)

Puzzles

- Why is there so little heterozygosity?
- Small animals have large populations and should have high heterozygosity. Why don't they?

Much of the rest of this course is about these questions.

A Python program to calculate $\hat{\mathcal{H}}$, using biallelic model

$$\hat{\mathcal{H}} = \frac{4Nu}{8Nu + 1} = \frac{\theta}{2\theta + 1}$$

```
# Expected heterozygosity as a function
# of theta = 4*N*u
def h(theta):
    return(theta/(2*theta + 1.0))

for theta in [0.001, 0.01, 0.1, 1.0, 10, 100]:
    print "%8.3f %8.3f" % (theta, h(theta))
```

Expected heterozygosity at a biallelic locus

$$\hat{\mathcal{H}} = \frac{\theta}{2\theta + 1}$$

θ	$\hat{\mathcal{H}}$
0.001	0.001
0.010	0.010
0.100	0.083
1.000	0.333
10.000	0.476
100.000	0.498

Expected heterozygosity at a biallelic locus